Responsiveness, Sensitivity and Clinical Utility of Timing-Related Speech Biomarkers for Remote Monitoring of ALS Disease Progression

Hardik Kothare[1]*, Michael Neumann*[1]*, Jackson Liscombe*[1]*, Jordan Green*[2]*, and Vikram Ramanarayanan*[1,3]

[1] Modality.AI, Inc., San Francisco, CA, USA
[2] MGH Institute of Healthcare Professions, Boston, MA USA
[3] University of California, San Francisco, CA, USA

<hardik.kothare, v>@modality.ai

Abstract

In this study, we describe the responsiveness of timing-related measures extracted from read speech in persons with ALS (pALS) collected via a remote patient monitoring platform in an effort to quantify how long it takes to detect a clinically-meaningful change associated with disease progression. We found that the timing alignment of pALS speech relative to a canonical elicitation of the same prompt is the most responsive measure, of the ones considered in this study, at detecting such change in both pALS with bulbar ($n = 35$) and non-bulbar onset ($n = 94$). We further evaluated the sensitivity of speech metrics in tracking disease progression in pALS while their ALSFRS-R speech score remained unchanged at 3 out of a total possible score of 4. We observed that timing-related speech metrics showed significant longitudinal changes even after accounting for learning effects. The findings of this study have the potential to inform disease prognosis and functional outcomes of clinical trials.

Index Terms: speech biomarkers, amyotrophic lateral sclerosis, remote patient monitoring

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive motor neuron disease with an estimated global prevalence of 4.42 per 100,000 persons [1]. The rapid degeneration and death of motor neurons lead to muscular atrophy, loss of voluntary motor control in persons with ALS (pALS) and a median survival of 3 to 5 years [2] after disease onset. Up to 30% of pALS present with bulbar onset of ALS, characterised by a rapid loss of speech and swallowing functions [3], while the rest present with non-bulbar onset characterised by muscular atrophy in the limbs and the trunk [4]. A vast majority of non-bulbar onset pALS eventually exhibit bulbar symptoms in the course of their disease progression [2]. Due to this heterogeneity in disease onset and progression, it is important to identify efficacious bulbar biomarkers to improve the predictive modelling of disease progression.

The current clinical gold standard to track disease progression in ALS is the ALS Functional Rating Scale - Revised (ALSFRS-R) [5], a questionnaire comprising 12 questions across four functional domains impacted by ALS [6]: bulbar, fine motor, gross motor and respiratory. However, there is evidence that the ALSFRS-R scale may track disease progression in a non-linear manner and may lack sensitivity in the early stages of bulbar disease onset [7, 8]. On the other hand, objective speech measures have been shown to be very powerful in early detection of bulbar symptoms [3, 8, 9, 10, 11, 12] and the progression of bulbar decline in pALS [13, 14, 15]. Specifically, Eshghi et al. [15] demonstrated that speaking rate and speech intelligibility can predict speech loss based on pre-defined thresholds and that these objective speech measures are more responsive to functional decline than patient-reported ALSFRS-R scores. Stegmann et al. [14] demonstrated that disease progression in bulbar onset and non-bulbar onset pALS can be predicted using speaking rate and articulatory precision using data collected remotely via a mobile application. Speaking rate has been consistently found to be an important biomarker for early diagnosis and stratification in both these studies and other studies, along with other timing-related measures like percentage pause time, speaking duration and others [10, 11, 16, 17]. Thus, timing-related speech biomarkers have the potential to track functional outcomes and slowing of bulbar decline in the context of clinical interventional trials targeting neurodegenerative disorders. To establish the efficacy of timing-related speech biomarkers in tracking bulbar decline, it is important to consider what constitutes a minimal clinically-important difference (MCID) [18, 19] instead of predefined thresholds that may vary by clinical phenotypes. For these speech biomarkers to be considered clinically useful, it is important that they are sensitive in detecting bulbar deterioration, which could be well before corresponding changes are observed in the relevant ALSFRS-R functional scores. To address the need for improved biomarkers of bulbar disease progression in ALS, we explored the responsiveness, sensitivity and clinical utility of four timing-related speech metrics by formulating the following research questions:

1. Is the rate of change in timing-related speech biomarkers different for bulbar and non-bulbar onset pALS? If so, can we quantify how different the rates are?
2. How many weeks does it take to detect a clinically meaningful change from disease onset using these metrics in both cohorts of pALS?
3. Can these metrics detect speech deterioration during intervals of time when patients report no speech changes (i.e., on the ALSFRS-R)?

2. Data

The study protocol was granted exempt status by an external Institutional Review Board[1]. Participants were recruited by EverythingALS and the Peter Cohen Foundation[2]. Data was collected in an ongoing study from 129 pALS (64 female, mean age $\pm$ standard deviation = 62.63 $\pm$ 7.98 years, Bulbar onset: n = 35, Non-Bulbar onset: n = 94) and 135 age and sex-matched controls (71 female, mean age $\pm$ standard deviation = 62.75 $\pm$ 8.06 years) between 2020-11-03 and 2023-02-07. For age matching, a tolerance threshold of ± 3 was set. Audiovisual data

[1] https://www.advarra.com/
[2] https://www.everythingals.org/research

from all participants was collected using the Modality platform, a cloud-based multimodal dialogue system in which participants perform standard motor speech tasks in a structured conversation with a virtual agent, Tina. Participants also filled out the ALSFRS-R survey after completing the standard battery of speech tasks. A total number of 2362 pALS sessions (506 bulbar onset and 1856 non-bulbar onset) and 3044 control sessions were considered in this analysis. On average, pALS completed 23 ($\pm$ 20) sessions, and controls completed 22 ($\pm$ 18) sessions.

3. Metrics

We focus on the Bamboo passage, a standardised 99-word reading passage designed to examine pausing behaviour by having voiced consonants at word and phrase boundaries. Timing-related measures during the reading of this passage are useful in detecting motor speech abnormalities in readers [10, 16, 17]. During every participant session, metrics were extracted automatically after segmentation of audio data collected at a sampling rate of 48kHz. We selected four timing-related metrics described below:

1. Speaking duration (seconds): The total amount of time taken to read the Bamboo passage in seconds.
2. Speaking rate (words per minute): The total number of words in the passage (99) divided by the time taken to read the Bamboo passage [20] .
3. Percentage pause time (PPT; %): The proportion of the total duration of all pauses to the total duration of the utterance.
4. Canonical Timing Alignment (CTA; %): A number between 0% (non-alignment) and 100% (perfect alignment) as measured by the normalised inverse Levenstein edit distance between words and silence boundaries. The participant's predicted word-level timing information, derived using the Montreal Forced Aligner [21], is compared to the expected production by Tina [22].

4. Methods

4.1. Clinically-meaningful change

To define a clinically meaningful change, we calculated the minimal clinically-important difference (MCID) [18, 19] for the four metrics described in this paper for a corresponding one-point change on the ALSFRS-R speech question where participants are asked to rate their speech on the following scale with scores in parentheses:

- Normal speech processes (4)
- Detectable speech disturbance (3)
- Intelligible with repeating (2)
- Speech combined with nonvocal communication (1)
- Loss of useful speech (0)

The MCID is the smallest domain-specific change that is thought to be clinically relevant [23]. It can be quantified as a threshold for a change corresponding to clinical improvement or deterioration [24] and is tied to an external anchor which is considered to be a clinical gold standard, the ALSFRS-R speech question in this case. The point representing maximum sensitivity and specificity (top left corner) on a receiver operating characteristic (ROC) curve of a simple binary classifier is the optimal cutpoint corresponding to the MCID value. MCID calculation was performed using the `rpy2` package in Python along with the `pROC` [25], `ROCR` [26] and `OptimalCutpoints` [27] packages in R [28]. The classes being discriminated in the classifier were pALS who exhibited a one-point change in their ALSFRS-R speech score and those who did not show a change in their ALSFRS-R speech score.

4.2. Longitudinal analysis

To evaluate the responsiveness and sensitivity of the metrics on a longitudinal basis, we used growth curve models (GCMs) [29] that provide a linear fit for a non-linear mixed effects model to estimate the trajectory of a metric over time with random slopes and intercepts for each participant [14]. The advantage of growth curve models is that they produce estimates of smoothed trajectories of change over time by using observed repeated measures of each individual, making it the ideal statistical method for the dataset considered in this paper. The assumption is that a latent growth process (decline of speech functions) is responsible for the change in observed measures. GCM fitting was performed in R. GCM curves for distinct cohorts can help identify differences in the longitudinal trajectory of measures in the two cohorts. In this paper's growth curve models, more than 90% of participants had at least 3 repeated measures, thus minimising any impact of variability in the number of sessions per participant [30].

4.2.1. Responsiveness

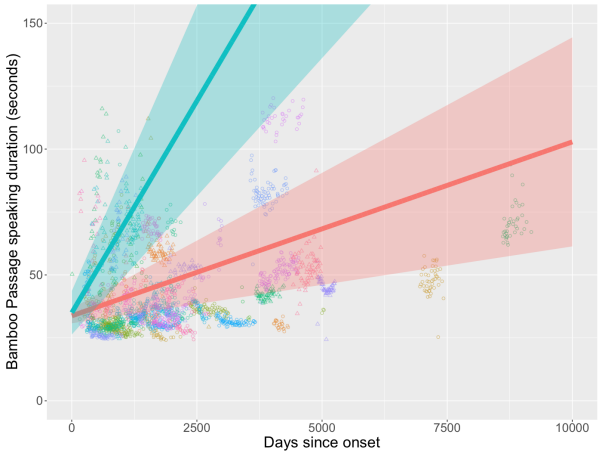
For the responsiveness analysis, the two cohorts chosen for growth curve modelling were sessions from pALS with bulbar onset and those from pALS with non-bulbar onset. Responsiveness was evaluated in two ways: (i) the time taken in weeks to detect deterioration greater than the standard error of the mean value for the cohort (statistical utility) and (ii) the time taken in weeks to detect deterioration greater than the MCID value (clinical utility).

4.2.2. Sensitivity

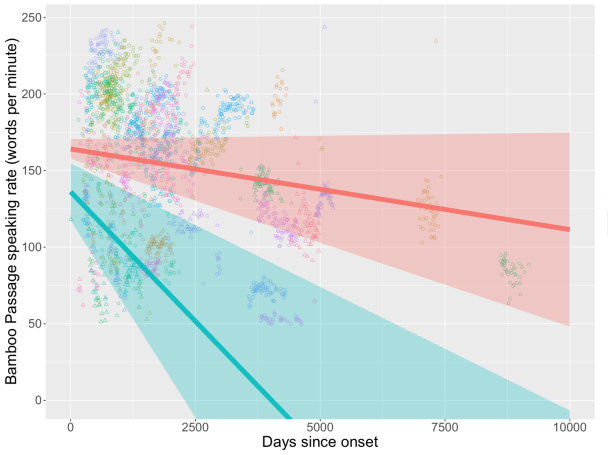
For sensitivity analysis, the two cohorts were sessions from healthy controls (where the ALSFRS-R speech score remained unchanged at 4: normal speech processes) and all contiguous pALS sessions with a speech score of 3. We decided to look at pALS sessions with a speech score of 3 because these pALS were deemed to exhibit bulbar impairment (albeit per self-perception) but still had speech that was intact enough for objective analysis. A metric was determined to be sensitive if the slope of the GCM for pALS with a steady speech score of 3 varied as compared to the slope of participants from the control cohort with a steady speech score of 4. Longitudinal data may be confounded by the presence of learning effects due to the repetition of the same tasks over time. In the case of the Bamboo passage, familiarity with the words in the passage may lead to an increased speaking rate. The advantage of comparing the trajectory of metrics in 'clinically-stable' pALS with that in controls is that it will demonstrate a difference in slopes over any learning effects (assuming the learning effects are equal across cohorts).

5. Results

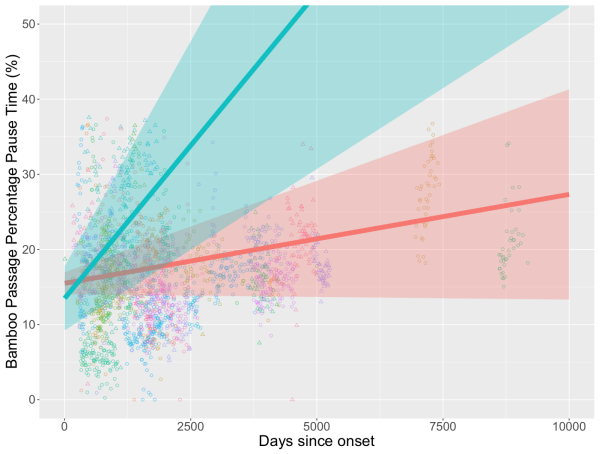
All four timing-related metrics showed differences in slopes between bulbar onset and non-bulbar onset pALS with the bulbar onset cohort exhibiting a steeper slope or a more rapid deterioration in speech (see Figure 1). Details of the slopes per cohort and responsiveness values can be found in Table 1. Speaking rate was found to be the measure with the most responsive statistical utility (2.97 weeks in pALS with bulbar onset). However, if both statistical and clinical utility are taken into account, CTA was found to be the most responsive measure in both cohorts. CTA shows statistical and clinical utility in detecting



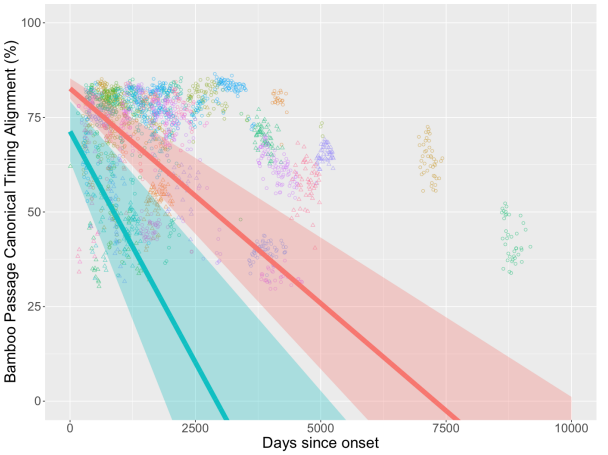
(a) *Speaking duration (p = 0.0007; equations: bulbar onset speaking duration = 0.2358 * number of weeks + 34.97; non-bulbar onset speaking duration = 0.0491 * number of weeks + 33.76)*



(b) *Speaking rate (p = 0.0159; equations: bulbar onset speaking rate = -0.2373 * number of weeks + 136.14; non-bulbar onset speaking rate = -0.0344 * number of weeks + 164.08)*



(c) *Percentage Pause Time (p = 0.0070; equations: bulbar onset PPT = 0.0568 * number of weeks + 13.48; non-bulbar onset PPT = 0.008 * number of weeks + 15.49)*



(d) *Canonical Timing Alignment (p = 0.0344; equations: bulbar onset CTA = -0.1712 * number of weeks + 71.31; non-bulbar onset CTA = -0.0793 * number of weeks + 82.64)*

Figure 1: *Growth curve models showing steeper rates of change for timing-related metrics in bulbar onset pALS (blue) as compared to non-bulbar onset pALS (red). Note: The cohort-specific lines in the growth curve model figures are not linear regression fits. They represent the average intercept and slope across all participants in the respective cohorts. Each data point represents a session.*

bulbar onset pALS within less than 4 weeks and in non-bulbar onset pALS within 9 weeks. Although speaking rate and PPT show differences in the longitudinal trajectory between bulbar onset and non-bulbar onset pALS, the time taken to observe a clinical change, especially in non-bulbar onset pALS, may be too long to be of clinical utility for some interventional trials.

All four metrics were also sensitive enough to show a longitudinal change before any change in the ALSFRS-R speech score of patients from 3 (see Table 2). Since clinical deterioration of speech in controls is not expected, any changes in metrics can be attributed to familiarisation with the task or learning effects. A learning effect in controls can be observed through the negative slope for speaking duration and PPT and a positive slope for speaking rate. Note that the slope for CTA is negative in controls because an increase in speaking rate would reduce the CTA value because the elicitation will be faster than the canonical elicitation of the reading passage. However, the longitudinal trajectory of CTA in controls had a slope that was not statistically different from 0, i.e. negligible. Differences between controls and pALS were observed despite the presence of these learning effects.

6. Discussion and Conclusions

In this work, we investigated whether the longitudinal trajectories of timing-related speech biomarkers extracted from read speech can distinguish between pALS with bulbar onset and those with non-bulbar onset. All four biomarkers — speaking duration, speaking rate, PPT and CTA — showed a steeper decline in pALS with bulbar onset associated with speech deterioration. Of these four metrics, CTA was the most responsive, i.e. per the growth curve models, it took the shortest time to detect a change that was statistically and clinically relevant. Previous studies have demonstrated the utility of speaking rate in distinguishing between longitudinal changes in bulbar onset and non-bulbar onset cohorts [14, 15]. In this study, since speaking rate was calculated as the total speaking duration divided by the total number of words, one would expect to see similar trajectories for both metrics. However, the assumption here is that ev-

Table 1: *Responsiveness of metrics. Bulbar onset: n = 35 (506 sessions), Non-Bulbar onset: n = 94 (1856 sessions)*

Metric	MCID	Onset	Slope ± standard error of slope per week	Standard error (SE) of the mean	Weeks to detect change > SE	Weeks to detect change > MCID
Speaking duration (seconds)	1.91	Bulbar	0.2358 ± 0.055	0.70	2.97	8.10
		Non-Bulbar	0.0491 ± 0.027	0.41	8.35	38.90
Speaking rate (words per minute)	-6.57	Bulbar	-0.2373 ± 0.0841	1.33	5.60	27.68
		Non-Bulbar	-0.0344 ± 0.04	1.073	31.19	190.99
PPT (% points)	3.92	Bulbar	0.0568 ± 0.0181	0.32	5.63	69.01
		Non-Bulbar	0.008 ± 0.0087	0.14	17.50	490.00
CTA (% points)	-0.66	Bulbar	-0.1712 ± 0.0403	0.53	3.10	3.86
		Non-Bulbar	-0.0793 ± 0.0206	0.33	4.16	8.32

Table 2: *Sensitivity of metrics. 38 pALS (684 sessions)*

Metric	p-value of difference	Cohort	Intercept ± standard error	Slope ± standard error
Speaking duration (seconds)	<0.0001	Controls	34.66 ± 1.03	-0.0431 ± 0.0133
		pALS	44.22 ± 2.32	0.1417 ± 0.0308
Speaking rate (words per minute)	<0.0001	Controls	173.90 ± 2.72	0.2507 ± 0.0337
		pALS	136.15 ± 6.12	-0.2474 ± 0.0778
PPT (% points)	0.0112	Controls	14.94 ± 0.57	-0.0148 ± 0.0079
		pALS	17.06 ± 1.30	0.0313 ± 0.0182
CTA (% points)	<0.0001	Controls	78.44 ± 0.96	-0.0076 ± 0.0133
		pALS	68.65 ± 2.17	-0.1628 ± 0.0294

ery participant finished reading the entire passage (99 words). Since the trajectories are slightly different, it would be fair to assume that some participants did not read the whole passage. Therefore, between the two metrics, speaking duration is probably a stronger measure because it does not assume task completion. In future work, automatic speech recognition will be used to ensure task completion. Relatedly, we also looked at differences in articulation rate (total duration of reading excluding pauses divided by total number of words) between the two cohorts. While pALS with bulbar onset and those with non-bulbar onset had statistically different intercepts, we did not observe differences in the slopes of the longitudinal trajectories of the metric. This is indicative of cohort-specific differences in pausing patterns rather than changes in articulation, thus contributing to differences in PPT and CTA over time. We also aimed to investigate whether the four timing-related metrics show a statistically significant change over time while the clinical gold standard indicated no clinical change in bulbar-impaired pALS. For this, we chose pALS who perceived their speech to be impaired, that is a score of 3 on the ALSFRS-R speech question. We hypothesise that pALS with a score of 4 would probably be the ones without any bulbar or speech motor impairment (at least at the time of measurement), and would show a ceiling effect for most metrics, which would be unsuitable for studying their sensitivity. In fact, when we looked at pALS with a steady speech score of 4, their longitudinal trajectory was not statistically different from controls. Unsurprisingly, controls showed a learning effect over time for speaking rate and speaking duration (a slope statistically different from 0) as they got more familiar with the Bamboo passage but not for PPT and CTA, indicating that pausing patterns in controls did not change over time. However, pALS displayed metric trajectories that were different from controls even when learning effects were present in controls indicating a high sensitivity for these metrics to detect speech deterioration faster than the subjective clinical survey instrument. One limitation here is the assumption that pALS and controls would exhibit similar rates of learning.

The analyses performed in this paper inherently assumed linearity of ALS disease progression. While this was done for simplicity and ease of interpretation and is still useful, we know that this assumption is not accurate. Research has shown that ALS progression is frequently nonlinear, with periods of stable disease preceded or followed by rapid decline [31]. Future work will focus on improving modelling methods to better capture trajectories and the variability in different clusters of patients that may share similar disease progression patterns. These timing-related metrics were calculated using data collected remotely and thus have the potential to be included in large interventional trials with geographically-distributed populations. Since CTA requires a maximum time of 8.32 weeks to detect clinical changes in both bulbar and non-bulbar onset pALS, it provides an opportunity to track speech change and even its slowing down in a relatively short period of time. To our knowledge, this is the first work to compare both the statistical and the clinical responsiveness of timing-related metrics in bulbar and non-bulbar onset ALS and to report changes in more finely grained objective metrics before any changes in the current clinical gold-standard survey instrument (ALSFRS-R) are reported.

In conclusion, the longitudinal trajectories of timing-related speech biomarkers associated with the reading of a passage are useful in distinguishing between persons with bulbar onset ALS and non-bulbar onset ALS. These trajectories help determine how many weeks it takes to detect clinically-important speech deterioration. Among the four biomarkers tested, the timing alignment of read speech as compared to a canonical reading of the passage was the most responsive to bulbar decline. Additionally, these biomarkers are sensitive enough to detect a change before any clinical change is detected by the prevalent gold-standard survey instrument, the ALSFRS-R scale.

7. Acknowledgements

This work was funded by the National Institutes of Health grant R42DC019877. We thank our collaborators at EverythingALS and the Peter Cohen Foundation for participant recruitment and data collection. We also thank Gabriela M. Stegmann for providing the template R code to fit growth curve models.

8. References

[1] L. Xu, T. Liu, L. Liu, X. Yao, L. Chen, D. Fan, S. Zhan, and S. Wang, "Global variation in prevalence and incidence of amyotrophic lateral sclerosis: A systematic review and meta-analysis," *Journal of Neurology*, vol. 267, pp. 944–953, 2020.

[2] L. J. Haverkamp, V. Appel, and S. H. Appel, "Natural history of amyotrophic lateral sclerosis in a database population validation of a scoring system and a model for survival prediction," *Brain*, vol. 118, no. 3, pp. 707–719, 1995.

[3] J. R. Green, Y. Yunusova, M. S. Kuruvilla, J. Wang, G. L. Pattee, L. Synhorst, L. Zinman, and J. D. Berry, "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, pp. 494–500, 2013.

[4] L. C. Wijesekera and P. Nigel Leigh, "Amyotrophic lateral sclerosis," *Orphanet Journal of Rare Diseases*, vol. 4, pp. 1–22, 2009.

[5] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. 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.

[6] K. Rascovsky, S. Xie, A. Boller, X. Han, L. McCluskey, L. Elman, and M. Grossman, "Subscales of the ALS functional rating scale (ALSFRS-R) as determinants of survival in amyotrophic lateral sclerosis (ALS)(p4. 094)," 2014.

[7] M. Proudfoot, A. Jones, K. Talbot, A. Al-Chalabi, and M. R. Turner, "The ALSFRS as an outcome measure in therapeutic trials and its relationship to symptom onset," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 17, no. 5-6, pp. 414–425, 2016.

[8] K. M. Allison, Y. Yunusova, T. F. Campbell, J. Wang, J. D. Berry, and J. R. Green, "The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 18, no. 5-6, pp. 358–366, 2017.

[9] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, "Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis," in *Proc. Interspeech 2018*, 2018, pp. 377–381.

[10] C. Barnett, J. R. Green, R. Marzouqah, K. L. Stipancic, J. D. Berry, L. Korngut, A. Genge, C. Shoesmith, H. Briemberg, A. Abrahao *et al.*, "Reliability and validity of speech & pause measures during passage reading in ALS," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 21, no. 1-2, pp. 42–50, 2020.

[11] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, D. Pautler, I. Navar, A. Anvar, J. Kumm, R. Norel, E. Fraenkel, A. V. Sherman, J. D. Berry, G. L. Pattee, J. Wang, J. R. Green, and V. Ramanarayanan, "Investigating the Utility of Multimodal Conversational Technology and Audiovisual Analytic Measures for the Assessment and Monitoring of Amyotrophic Lateral Sclerosis at Scale," in *Proc. Interspeech 2021*, 2021, pp. 4783–4787.

[12] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green, "Speech as a biomarker: Opportunities, interpretability, and challenges," *Perspectives of the ASHA Special Interest Groups*, pp. 1–8, 2022.

[13] P. Rong, Y. Yunusova, and J. R. Green, "Speech intelligibility decline in individuals with fast and slow rates of ALS progression," in *Proc. Interspeech 2015*, 2015, pp. 2967–2971.

[14] G. M. Stegmann, S. Hahn, J. Liss, J. Shefner, S. Rutkove, K. Shelton, C. J. Duncan, and V. Berisha, "Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis," *NPJ Digital Medicine*, vol. 3, no. 1, p. 132, 2020.

[15] M. Eshghi, Y. Yunusova, K. P. Connaghan, B. J. Perry, M. F. Maffei, J. D. Berry, L. Zinman, S. Kalra, L. Korngut, A. Genge *et al.*, "Rate of speech decline in individuals with amyotrophic lateral sclerosis," *Scientific Reports*, vol. 12, no. 1, p. 15713, 2022.

[16] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," *Journal of Medical Speech-Language Pathology*, vol. 12, no. 4, p. 149, 2004.

[17] Y. Yunusova, N. L. Graham, S. Shellikeri, K. Phuong, M. Kulkarni, E. Rochon, D. F. Tang-Wai, T. W. Chow, S. E. Black, L. H. Zinman *et al.*, "Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)," *PLOS ONE*, vol. 11, no. 1, p. e0147573, 2016.

[18] A. E. McGlothlin and R. J. Lewis, "Minimal clinically important difference: defining what really matters to patients," *JAMA*, vol. 312, no. 13, pp. 1342–1343, 2014.

[19] K. L. Stipancic, Y. Yunusova, J. D. Berry, and J. R. Green, "Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 61, no. 11, pp. 2757–2771, 2018.

[20] K. Yorkston, D. Beukelman, and R. Tice, "Sentence intelligibility test," *Lincoln, NE: Tice Technologies*, 1996.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[22] J. Liscombe, M. Neumann, H. Kothare, O. Roesler, D. Suendermann-Oeft, and V. Ramanarayanan, "On timing and pronunciation metrics for intelligibility assessment in pathological ALS speech," in *Vol 27: Suppl. (2022): Abstracts 8th International Conference on Speech Motor Control Groningen, August 2022*, 2022.

[23] H. C. de Vet, C. B. Terwee, R. W. Ostelo, H. Beckerman, D. L. Knol, and L. M. Bouter, "Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change," *Health and Quality of Life Outcomes*, vol. 4, pp. 1–5, 2006.

[24] H. Kothare, M. Neumann, J. Liscombe, O. Roesler, W. Burke, A. Exner, S. Snyder, A. Cornish, D. Habberstad, D. Pautler *et al.*, "Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for Parkinson's disease assessment," *Proc. Interspeech 2022*, pp. 3658–3662, 2022.

[25] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, vol. 12, p. 77, 2011.

[26] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, p. 7881, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de

[27] M. López-Ratón, M. X. Rodríguez-Álvarez, C. C. Suárez, and F. G. Sampedro, "OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests," *Journal of Statistical Software*, vol. 61, no. 8, pp. 1–36, 2014.

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[29] D. Von Rosen, "The growth curve model: a review," *Communications in Statistics-Theory and Methods*, vol. 20, no. 9, pp. 2791–2822, 1991.

[30] P. J. Curran, K. Obeidat, and D. Losardo, "Twelve frequently asked questions about growth curve modeling," *Journal of Cognition and Development*, vol. 11, no. 2, pp. 121–136, 2010.

[31] D. Ramamoorthy, K. Severson, S. Ghosh, K. Sachs, J. D. Glass, C. Fournier, A. Sherman, T. M. Herrington, J. Berry, and E. Fraenkel, "Identifying patterns in amyotrophic lateral sclerosis progression from sparse longitudinal data," *Nature Computational Science*, vol. 2, no. 9, pp. 605–616, 2022.