Multimodal Digital Biomarkers for Longitudinal Tracking of Speech Impairment Severity in ALS: An Investigation of Clinically Important Differences

Hardik Kothare[1*], *Michael Neumann*[1*], *Jackson Liscombe*[1], *Emma C.L. Leschly*[1], *Oliver Roesler*[1], *Vikram Ramanarayanan*[1,2]

[1]Modality.AI, Inc., San Francisco, CA, USA
[2]University of California, San Francisco, San Francisco, CA, USA

v@modality.ai

Abstract

Speech biomarkers have shown promise for the remote assessment of ALS. However, to demonstrate clinical utility at tracking longitudinal progress of the disease, one needs to understand how well these biomarkers capture changes that are 'clinically meaningful', a concept that is not always clearly defined. Therefore, this paper defines and explores multiple methods of computing minimal clinically important difference (MCID) using ratings of speech impairment severity and listener effort as clinical anchors. We analyze how these methods impact the estimated responsiveness of various metrics collected from 125 ALS patients via a multimodal dialog based remote assessment platform. We find that select biomarkers are more responsive than the clinical standard ALSFRS-R across the board at tracking clinically meaningful changes related to speech severity. We further discuss advantages and disadvantages of different MCID computation methods for assessing ALS disease progression.

Index Terms: speech biomarkers, multimodal, dialog, ALS, remote assessment, minimal clinically important difference, speech impairment, listener effort

1. Introduction

Speech-based digital biomarkers have shown excellent potential for assessment of neurological conditions like Amyotrophic Lateral Sclerosis, or ALS [1, 2, 3, 4]. However, before these can be operationally deployed in the real world for remote patient monitoring in clinical care or as endpoints in clinical trials, one needs to demonstrate clinical utility of these biomarkers at tracking longitudinal progress of a disease like ALS. And in order to do that, one needs to understand how well these biomarkers capture changes that are 'clinically meaningful'.

The concept of minimal clinically important difference (MCID) is an attempt to capture change in a biomarker that is clinically meaningful. It quantifies the minimal change in a clinical outcome measure that is considered relevant or important for patients, caregivers and/or clinicians [5, 6]. It is well established for measuring improvement after a treatment, often defined by patient reported outcomes (PRO) such as questionnaire based scales. Here, the MCID is the smallest change that is considered meaningful and worthwhile by the patient to undergo a particular treatment. In the same manner, MCID can be utilized to estimate the minimal threshold for deterioration in symptoms to be considered important for patients, and can therefore be used to quantify disease progression. To determine the MCID, two general approaches can be used, anchor-based and distribution-based methods. To date, there seems to be no consensus on using one specific method, instead, it is advised to use a combination of methods to narrow down a range of MCID thresholds [6]. While distribution-based methods can be used without external information based on the outcome measure's statistical properties (e.g., by defining the MCID as a fraction of the standard deviation), they lack the *clinical meaningfulness*. It is generally preferred to use anchor-based methods that tie changes in the outcome to an external source of clinical relevance, e.g., to a validated questionnaire based scale. For speech measures in ALS, Stipancic et al. [7] presented such an approach based on changes in the ALS functional rating scale - revised (ALSFRS-R), the standard instrument to assess progression in ALS [8]. Following this approach, Kothare et al. investigated the responsiveness and sensitivity of speech biomarkers to track change in ALS [9].

These previous studies have relied on the ALSFRS-R speech sub score as the external anchor, which limits the granularity at which meaningful changes can be identified because speech impairment is rated using only one question on a five-point scale (ranging from "normal speech processes" to "loss of useful speech"). Further, only one method to derive an MCID was explored in these studies, which was based on receiver operating characteristics (ROC) analysis. For a more sensitive external anchor with respect to speech impairment in ALS, we propose to use a visual analog scale (VAS) to rate listener effort. This approach has been shown to produce reliable ratings of impairment severity [10, 11]. Instead of limiting the MCID estimation to one method, we investigate multiple approaches, resulting in a range of MCID thresholds rather than a single value for each speech feature. This paper attempts to answer the following research questions: (1) using VAS ratings of speech impairment severity and listener effort as external anchors to estimate MCID, how responsive (in terms of the estimated time it takes to detect meaningful change) are multimodal speech biomarkers computed from a remote ALS monitoring platform, as compared to the clinical standard ALSFRS-R? (2) what are the relative advantages and disadvantages of different MCID estimation methods and clinical anchors, and what do these mean for ALS clinical trials and progress monitoring?

2. Data

Audiovisual data from 143 people with ALS, or pALS (70 female; 36 with bulbar onset; mean age $\pm$ standard deviation = 60.4 ± 10.2) were recorded between November 2020 and February 2024 using a web based multimodal dialog platform [12, 13, 14]. The ongoing data collection was granted exempt status by an external Institutional Review Board.[1] Participants were recruited by EverythingALS and the Peter Cohen Foun-

*Both authors contributed equally.

[1]https://www.advarra.com/

dation[2]. The speech assessment contained a number of standard tasks, which have been adapted to the self-guided remote setting, including among others a diadochokinesis test (DDK), sentence intelligibility test (SIT), a reading passage (RP; Bamboo passage, 99 words), and a picture description (PD) task. After completing the speech assessment, participants filled out the ALSFRS-R. The dataset contains 3,350 sessions, the mean number of sessions per particpant is 23.4 ($\pm$ 24.4), and the mean duration between first and last assessment is 12.1 months ($\pm$ 10.8 months). 18 particpants had only one session and were excluded from further analysis.

3. Methods

3.1. Feature Extraction

Speech metrics were automatically extracted from the audio recordings using Praat [15] and the Montreal Forced Aligner [16]. Speech metrics included, among others, fundamental frequency (F0), harmonics-to-noise ratio (HNR), cepstral peak prominence (CPP), duration and pausing measures, canonical timing alignment (CTA)[3] [17], speaking rate, jitter, and shimmer. Facial video metrics, such as kinematics of articulators (jaw, lower lip), surface area of the mouth, and eyebrow raises were derived from facial landmarks generated with MediaPipe Face Mesh [18]. These metrics were normalised by dividing their values by the inter-caruncular distance [19]. Linguistic metrics were extracted from automatic transcriptions[4] of the picture description task, using the Python package `spaCy` [20].

To identify representative features from a large set of multicollinear features, we used hierarchical clustering on the Spearman rank-order correlations [21]. For this feature selection step, data from 135 healthy controls (71 female; mean age (standard deviation) = 59.9 (10.3) years) was used. Ward's method was used for clustering and feature clusters were visually inspected from a dendrogram. A distance threshold was chosen manually to divide clusters that represent sensible feature groupings in terms of the domain, resulting in 27 clusters. One representative feature within each cluster was selected by performing ROC analysis to determine the area under the ROC curve (AUC) for distinguishing bulbar onset participants from non-bulbar onset participants. To further filter features, we imposed a minimum threshold for the ROC-AUC. Features with an AUC$\geq$ 0.65 and with significantly different longitudinal trajectories were selected for further analysis.

3.2. Listener Effort Ratings

Previous work has shown that perceptual ratings of listener effort align well with clinician severity ratings of speech impairment [10, 11]. We followed this approach to obtain listener effort ratings for 369 samples of the Bamboo reading passage. Using a visual analog scale, three human listeners (two speech scientists and one computer science student) and one clinically-trained speech language pathologist (SLP) provided a rating of how effortful it was to understand each speech sample. The VAS was presented on screen as a vertical line, 400 pixels in height, and had the labels "Not at all" and "Very" at the end points following work by Picou et al. [22] and Stipanic et al. [11]. Raters were asked, "How effortful was it for you to understand?". The position of the slider was converted into an integer score between 0 (not at all) and 100 (very). To reduce rating time, only the first 15 seconds were played for every sample. The samples were selected as follows: For each participant, their first and last assessment plus a third sample that was closest in time to the midpoint of the interval between the first and last session were selected.[5] Inter-rater agreement was assessed by means of the intra-class correlation coefficient (ICC).

3.3. Minimal Clinically Important Difference

To provide inclusive MCID estimates, we included both distribution-based and anchor-based methods. Anchor-based methods should be preferred if a meaningful clinical anchor is available, and distribution-based methods are useful to provide supportive evidence for the MCID [23]. The listener effort ratings serve as an external anchor of meaningfulness, assuming that a certain amount of change in the ratings reflects a relevant change in speech impairment, which directly affects quality of life. Because the perceptual ratings are continuous, the minimal important difference needs to be defined, for example based on expert judgement or based on statistical characteristics of the change in ratings from one time point to another.

Here, we defined that a minimal important difference of listener effort ratings between two assessments lies between the mean absolute deviation (MAD) of the mean change across all participants and twice the MAD to find all speech sample pairs with at least a statistically meaningful minimal change, but also leave out sample pairs where the magnitude of change is significantly higher. MAD was preferred over standard deviation (SD) because differences in ratings were not normally distributed. Participants were stratified into two groups: those whose ratings yielded a change between MAD and 2MAD and those whose ratings did not change (difference smaller than MAD). The difference in ratings was calculated between adjacent sessions. To compare this approach to the ALSFRS-R as external anchor, we followed [7] and utilized a 1-point decline in the ALSFRS-R speech score as minimal important difference. Analogously, participants were split into two groups, those with a 1-point decline and those with no change between sessions.

Having defined these two groups, three methods commonly found in the MCID literature [6] were used to derive the MCID estimate: (a) Change difference (CD): MCID = difference between the mean metric change in the first group and the mean metric change in the second group, (b) Average change (AC): MCID = average of the metric change in participants classified as having experienced change (first group)[6], and (c) ROC analysis: MCID = the threshold on the ROC curve that maximizes sensitivity and specificity for discriminating the two groups. For the ROC approach, we identified the point closest to the top left corner of the ROC curve using the `coords` function from the `pROC` R package [24]. The following distribution-based methods were used to estimate MCID without external anchor: (d) the standard error of measurement (SEM) of the change, defined as $SEM = SD\sqrt{1-r}$, where r is a measure of reliability (test-retest reliability of the feature in terms of Pearson correlation in our case), and (e) half the standard deviation (0.5SD) of the mean change. Both measures have been suggested in the literature [5, 6].

3.4. Longitudinal Analysis

To model longitudinal responsiveness of speech features, we fitted growth curve models (GCM) following the approach described in [25]. GCMs allow us to estimate a smoothed tra-

[2] `https://www.everythingals.org/research`

[3] CTA is a measure of word-level alignment between the spoken utterance and a canonical speech production of the same text.

[4] `https://aws.amazon.com/transcribe/`

[5] Six participants had only two sessions, so the final number of samples was $119 * 3 + 6 * 2 = 369$.

[6] If the mean change in the second group is 0, CD and AC are equal.

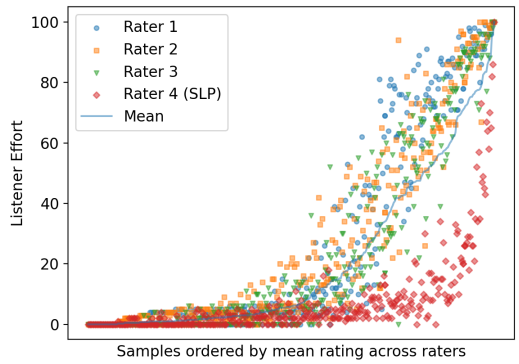

Figure 1: *Listener effort ratings across four raters, sorted by their mean rating, to illustrate the agreement among raters.*

jectory of a speech feature over time with random slopes and intercepts for each participant. For this analysis, the full dataset was used to provide as many data points for the model fit as possible. Because site of onset plays an important role in ALS progression, growth curve modelling was done separately for pALS with bulbar onset and pALS with non-bulbar onset [9]. The GCM trajectories were then used to calculate responsivenessof each metric as the average time in weeks that it takes to detect change in that metric greater than the MCID.

4. Results

4.1. Listener Effort Ratings

Fig. 1 shows the distribution of VAS ratings for listener effort across all four raters. It is evident that the SLP ratings differ significantly from the judgement of the other three raters. The reasons for this difference were examined by means of a group discussion among all raters. It turned out that raters 1, 2 and 3 had a different concept of listener effort in mind; they judged *perceivable speech disturbances* as being more effortful to understand, even if this did not translate to increased cognitive load. Instead, the SLP ratings were solely based on the notion of how much cognitive and mental effort is required to understand a speech sample. Indeed, previous work has shown differences in judgments of speech severity and listener effort between SLPs and other types of listeners [26]. Despite this difference, the trend of all ratings is similar, as confirmed by an excellent inter-rater agreement of ICC=0.92 ($p < 0.0001$, 95%-CI: [0.91, 0.94]). In fact, the VAS ratings of rater 1, 2 and 3 suggest more gradually increasing speech impairment severity in the data in terms of perceivable changes. We therefore decided to use the SLP's ratings separately as representative of listener effort, and the mean of the remaining three raters' scores as representative of *impairment severity*, labeled as *severity rating* for the rest of this paper.

	ρ	r	$\mathcal{N}$
RP speaking duration	0.79	0.80	354
SIT PPT	0.59	0.60	365
RP CPP	0.51	0.53	363
DDK HNR	0.35	0.41	363
SIT HNR	0.37	0.35	367
RP mean F0	0.17	0.23	364
RP max. lip width	0.15	0.19	361
ALSFRS-R total	-0.25	-0.16	270
PD word count	-0.63	-0.65	251
ALSFRS-R bulbar	-0.78	-0.77	277
ALSFRS-R speech	-0.80	-0.79	277
RP CTA	-0.74	-0.83	333

Table 1: *Correlation (Spearman's ρ and Pearson's r; all significant at $p < 0.01$) between the mean severity rating and metrics as well as ALSFRS-R scores. $\mathcal{N}$: sample size.*

Table 1 shows the correlation between the mean severity ratings and the selected features and ALSFSR-R scores in terms of Spearman and Pearson correlation coefficients. Timing-related metrics, such as duration, percent pause time (PPT), and canonical timing alignment (CTA) show strong correlation with dysarthria severity, and voice quality features, such as CPP and HNR show moderate correlation. The ALSFRS-R speech and bulbar sub scores are strongly correlated with severity ratings. However, we observed large overlaps in the 5-point distribution of the ALSFRS-R speech score with respect to severity ratings. Speech scores of 2 and 3 appear within a broad and largely overlapping range of severity scores, indicating the lack of sensitivity to track gradual changes. In contrast, metrics such as CTA showed a stronger linear relationship to dysarthria severity.

4.2. MCID Estimates

When severity ratings and listener effort were used as anchors, participants were stratified into two groups based on the MAD of the mean change in ratings. For severity ratings the MAD was 7.20 and the number of samples (adjacent sessions) was 28 in the change group and 160 in the unchanged group. For listener effort ratings the MAD was 5.49 for and the number of samples was 19 and 185, respectively. When taking a 1-point decline in the ALSFRS-R speech score as external anchor, the number of samples was 17 in the change group and 122 in the unchanged group. In addition, we calculated the MCID estimates for the ALSFRS-R speech anchor on the full dataset, taking adjacent sessions with an interval of at least 14 days to determine change in outcomes. For this, the number of samples was 57 in the change group and 1,355 in the unchanged group.[7]

Table 2 presents all MCID estimates. When we consider distribution-based methods, a 0.5 standard deviation (0.5SD) change provides a more conservative estimate (larger MCID) compared to the SEM for most features, potentially attributable to the good test-retest reliability values of these features[8]. No clear trends were observed regarding the relationship between distribution-based and anchor-based estimates, which underlines that statistical properties alone cannot serve as a reliable indicator of clinically important change.

Before comparing different choices of external anchors, we consider the three selected methods of calculating MCID based on having a change group and an unchanged group. All ROC analyses yielded consistently low area under the curve (AUC) values of below 0.7 (and in some cases even below 0.6) for most constellations, indicating relatively low discriminative ability. This is likely caused by the highly heterogeneous presentation of ALS in terms of disease progression. As a result, the ROC-derived MCIDs should be utilized with caution and put into context by exploring additional methods. In the present study, ROC-derived estimates were not considered for further analysis and discussion. With regard to the external anchor of clinical meaningfulness, the analysis revealed no clear trends across different metrics. Based on the dataset that was rated for severity and listener effort, the MCID estimates (AC and CD) are mostly within a small range for a given metric. Exception are higher absolute valus for CTA based on listener effort, and a large spread of estimates for word count (PD), which might be attributed to a large standard deviation in the metric itself.

[7]For pALS in the change group, only those adjacent sessions were taken into account where change was observed.

[8]The SEM equals 0.5SD for reliability r=0.75, and it is smaller for more reliable measures with r>0.75

Anchor			Severity rating			Listener effort			ALSFRS-R speech			ALSFRS-R speech*		
Cutpoint			MAD<change<2MAD			MAD<change<2MAD			1 point decline			1 point decline		
Method	0.5SD	SEM	CD	AC	ROC	CD	AC	ROC	CD	AC	ROC	CD	AC	ROC
RP CTA	3.16	1.69	-4.06	-4.91	$-2.66^{\ddagger}$	-5.42	-7.36	-3.85	-2.15	-4.86	$-3.67^{\ddagger}$	-1.89	-1.99	$-1.26^{\dagger}$
RP dur.	4.37	2.01	5.75	6.71	$3.44^{\ddagger}$	7.22	8.05	1.74	5.62	8.61	$7.38^{\ddagger}$	4.25	4.39	$1.45^{\ddagger}$
RP mean F0	7.88	3.86	0.46	2.43	$0.31^{\dagger}$	1.44	3.58	$-2.81^{\dagger}$	4.58	5.50	$7.27^{\ddagger}$	1.72	1.84	$4.50^{\dagger}$
RP CPP	1.40	1.50	-0.66	-0.45	$-0.29^{\dagger}$	0.85	1.01	$1.68^{\ddagger}$	0.92	1.07	$2.34^{\ddagger}$	0.05	0.10	$0.41^{\dagger}$
SIT PPT	1.97	1.89	1.54	1.85	$1.75^{\ddagger}$	-0.13	0.25	$-0.24^{\dagger}$	1.77	2.51	$0.09^{\dagger}$	0.82	0.90	$1.21^{\dagger}$
SIT HNR	0.93	0.67	0.95	0.91	$0.20^{\ddagger}$	0.66	0.83	$0.78^{\dagger}$	1.33	1.36	$0.52^{\ddagger}$	-0.22	-0.19	$-0.45^{\dagger}$
DDK HNR	1.22	1.03	1.53	1.60	$0.96^{\ddagger}$	0.95	1.16	$0.12^{\ddagger}$	1.07	1.07	$0.44^{\dagger}$	-0.14	-0.08	$-0.28^{\dagger}$
PD #words	20.78	21.64	-13.06	-8.91	$-0.50^{\dagger}$	-2.40	-2.42	$-2.50^{\dagger}$	-23.40	-19.57	$-5.50^{\dagger}$	-7.73	-7.44	$-2.50^{\dagger}$
RP max. LW	0.0602	0.0609	0.0062	0.0079	$0.0170^{\dagger}$	0.0222	0.0267	$0.0003^{\dagger}$	0.0276	0.0262	$0.0292^{\dagger}$	0.0211	0.0201	$0.0133^{\dagger}$
FRS-R speech	0.17	-	-0.12	-0.19	$-0.50^{\dagger}$	-0.12	-0.20	$0.50^{\dagger}$	-	-	-	-	-	-
FRS-R bulbar	0.49	-	-0.07	-0.25	$-0.50^{\dagger}$	-0.75	-1.0	$-0.50^{\ddagger}$	-	-	-	-	-	-

Table 2: *MCID estimates. For anchor-based methods, the sign indicates the direction of change. The distribution-based MCIDs are positive by definition. SD: standard deviation, CD: change difference, AC: average change, ROC: receiver operating characteristics, SEM: standard error of measurement, RP: reading passage, PD: picture description, LW: lip width. †ROC-AUC<0.6, ‡ROC-AUC<0.7, *the three rightmost columns present MCID estimated based on the full dataset, with shorter time intervals between sessions.*

	Onset	Slope per week ($\pm$ standard error)	Median MCID	Weeks until MCID min.	median
RP CTA	B	-0.1978 ± 0.0394	-3.11	8.5	15.7
(%)	NB	-0.0733 ± 0.0177		17.2	42.4
RP dur.	B	0.3228 ± 0.0652	5.68	6.2	17.6
(s)	NB	0.0647 ± 0.0308		22.4	87.8
RP mean	B	0.1232 ± 0.0400	3.00	3.8	24.4
F0 (Hz)	NB	0.0635 ± 0.0164		5.0	47.3
RP CPP	B	0.0101 ± 0.0050	0.89	5.2	87.7
(dB)	NB	0.002 ± 0.0022		26.5	442.9
SIT PPT	B	0.0371 ± 0.0082	1.66	3.5	44.6
(%)	NB	0.0041 ± 0.0036		21.8	403.9
SIT HNR	B	0.0179 ± 0.0055	0.87	10.6	48.6
(dB)	NB	0.0025 ± 0.0025		75.9	347.7
DDK HNR	B	0.0248 ± 0.0064	1.07	3.3	43.3
(dB)	NB	0.0033 ± 0.0028		24.4	325.2
PD #words	B	-0.1669 ± 0.0761	-7.59	14.4	45.5
	NB	0.0824 ± 0.0323		6.1	92.1
RP max.	B	0.0002 ± 0.0002	0.02	31.2	121.0
lip width	NB	-0.0001 ± 0.0001		3.2	242.0
FRS-R	B	-0.0096 ± 0.0033	-	104.2*	
speech	NB	-0.0049 ± 0.0016		204.1*	
FRS-R	B	-0.0263 ± 0.0088	-	38.0*	
bulbar	NB	-0.0155 ± 0.0043		64.5*	

Table 3: *Responsiveness of metrics as determined by GCMs. Number of weeks to detect change is based on the minimum/median absolute MCID across methods (except ROC). B: bulbar, NB: non-bulbar. *Number of weeks for FRS-R subscores is based on a 1-point change as the smallest measurable unit.*

MCID estimates for ALSFRS-R subscores in Table 2 are fractional, suggesting that a relevant change as measured by perceptual ratings occurs before it can be detected on the ALSFRS-R. This is because the scale does not allow for reporting fractional changes as designed; 1 point is the minimum possible change. Thus, the MCIDs are theoretical values and not practically actionable. When including the full dataset (right section in Table 2), some MCID estimates are notably smaller (absolute) than based on the smaller subset of the data. This indicates that the time interval between observations affects the MCID estimation (at least when using the ALSFRS-R speech anchor).

4.3. Tracking Longitudinal Progression

Table 3 presents the average slopes for all metrics and the ALSFRS-R sub scores, which were determined by fitting GCMs, and the time it takes to detect a change greater than the MCID. For this, we used the minimum and median MCID values across all methods (excluding ROC-derived thresholds), in order to report the best-case scenario and more realistic estimates. For the bulbar onset cohort, all metrics except max. lip width show on average a clinically important difference in a shorter amount of time than the ALSFRS-R speech score. The same is true for CTA, speaking duration and the mean F0 for the reading passage, and word count for picture description in the non-bulbar onset cohort. This suggests that remotely extracted speech biomarkers are significantly more responsive, and therefore more clinically meaningful, than the current clinical standard ALSFRS-R at tracking speech impairment severity.

5. Discussion

We found that visual analog scale (VAS) ratings of speech impairment severity and listener effort can serve as external anchors to estimate MCID of multimodal speech biomarkers computed from a remote ALS monitoring platform. Based on these perceptual ratings, we showed that theoretical MCID estimates for the clinical standard ALSFRS-R are lower than the minimal detectable difference of 1 point, and that select speech biomarkers are significantly more responsive (in terms of the estimated time it takes to detect meaningful change) as compared to the ALSFRS-R. This is encouraging evidence supporting the use of remotely-collected speech biomarkers for monitoring ALS progression. However, we observed a large variation of the MCID thresholds, depending on the anchor and the method. These choices have implications for biomarker/endpoint selection in clinical trials. While we attempted to come up with a good estimate of clinically meaningful change by using the median of several MCID estimates, the question of which MCID estimation method is the best remains an open one for future research.

There are important limitations to note. For instance, the average change (AC) estimate does not take the unchanged group into account and is usually larger (absolute values) than the change difference (CD) [6], which is true in most cases for the presented features. If the CD is larger than the AC in absolute terms, it indicates a change in the opposite direction in the unchanged group. We observed this most prominently for PD word count, where we hypothesize a training effect in certain cohorts with increasing word count over time. Furthermore, irrespective of the method used, for PPT, HNR, and CPP, we observed changes in sign for some MCID thresholds depending on the anchor or dataset. One possible explanation is that in all these cases the MCID estimates are so close to zero that they in fact do not indicate true change, which in turn suggests that the two groups do not exhibit significant change relevant to the used anchors in these metrics (on average). Further research on large, diverse datasets is required to address these limitations. To conclude, we provided a range of MCID estimates for a set of speech features relevant for ALS and reported different scenarios for the time it needs to detect meaningful change.

6. Acknowledgements

This work was funded by the National Institutes of Health grant R42DC019877. We thank the participants of this study for their time and EverythingALS, and the Peter Cohen Foundation for participant recruitment.

7. References

[1] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[2] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: the use of vocal biomarkers from research to clinical practice," *Digital biomarkers*, vol. 5, no. 1, pp. 78–88, 2021.

[3] H. P. Rowe, S. Shellikeri, Y. Yunusova, K. V. Chenausky, and J. R. Green, "Quantifying articulatory impairments in neurodegenerative motor diseases: A scoping review and meta-analysis of interpretable acoustic features," *International Journal of Speech-Language Pathology*, pp. 1–14, 2022.

[4] 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*, vol. 7, no. 1, pp. 276–283, 2022.

[5] A. G. Copay, B. R. Subach, S. D. Glassman, D. W. Polly, and T. C. Schuler, "Understanding the minimum clinically important difference: a review of concepts and methods," *The Spine Journal*, vol. 7, no. 5, pp. 541–546, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1529943007000526

[6] Y. Mouelhi, E. Jouve, C. Castelli, and et al., "How is the minimal clinically important difference established in health-related quality of life instruments? review of anchors and methods," *Health Qual Life Outcomes*, vol. 18, no. 1, p. 136, 2020. [Online]. Available: https://doi.org/10.1186/s12955-020-01344-w

[7] 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.

[8] 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," *Journal of Neurology Sciences*, vol. 169, no. 1-2, pp. 13–21, 1999.

[9] H. Kothare, M. Neumann, J. Liscombe, J. Green, and V. Ramanarayanan, "Responsiveness, Sensitivity and Clinical Utility of Timing-Related Speech Biomarkers for Remote Monitoring of ALS Disease Progression," in *Proc. INTERSPEECH 2023*, 2023, pp. 2323–2327.

[10] J. E. Sussman and K. Tjaden, "Perceptual Measures of Speech From Individuals With Parkinson's Disease and Multiple Sclerosis: Intelligibility and Beyond," *Journal of Speech, Language, and Hearing Research*, vol. 55, no. 4, pp. 1208–1219, 2012.

[11] K. L. Stipancic, K. M. Palmer, H. P. Rowe, Y. Yunusova, J. D. Berry, and J. R. Green, ""you say severe, i say mild": Toward an empirical classification of dysarthria severity," *Journal of Speech, Language, and Hearing Research*, vol. 64, no. 12, pp. 4718–4735, 2021.

[12] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill *et al.*, "NEMSI: A Multimodal Dialog System for Screening of Neurological or Mental Conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.

[13] V. Ramanarayanan, D. Pautler, L. Arbatti, A. Hosamath, M. Neumann, H. Kothare, O. Roesler, J. Liscombe, A. Cornish, D. Habberstad, V. Richter, D. Fox, D. Suendermann-Oeft, and I. Shoulson, "When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do," in *Proc. Interspeech*, 2023, pp. 678–679.

[14] V. Ramanarayanan, "Multimodal technologies for remote assessment of neurological and mental health," *Journal of Speech, Language, and Hearing Research*, pp. 1–8, 2024.

[15] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[16] 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.

[17] 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.

[18] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann, "Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs," *CoRR*, vol. abs/1907.06724, 2019. [Online]. Available: http://arxiv.org/abs/1907.06724

[19] O. Roesler, H. Kothare, W. Burke, M. Neumann, J. Liscombe, A. Cornish, D. Habberstad, D. Pautler, D. Suendermann-Oeft, and V. Ramanarayanan, "Exploring Facial Metric Normalization For Within- and Between-Subject Comparisons in a Multimodal Health Monitoring Agent," in *Companion Publication of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22 Companion. New York, NY, USA: Association for Computing Machinery, 2022, p. 160–165. [Online]. Available: https://doi.org/10.1145/3536220.3558071

[20] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[21] D. Ienco and R. Meo, "Exploration and Reduction of the Feature Space by Hierarchical Clustering," in *Proceedings of the 2008 Siam International Conference on Data Mining*. SIAM, 2008, pp. 577–587.

[22] E. Picou, T. Moore, and T. Ricketts, "The effects of directional processing on objective and subjective listening effort," *Journal of Speech Language and Hearing Research*, vol. 60, p. 199, 01 2017.

[23] D. Revicki, R. D. Hays, D. Cella, and J. Sloan, "Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes," *Journal of clinical epidemiology*, vol. 61, no. 2, pp. 102–109, 2008.

[24] 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.

[25] 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, pp. 1–5, 2020.

[26] P. Dagenais, C. Watts, L. Turnage, and S. Kennedy, "Intelligibility and acceptability of moderately dysarthric speech by three types of listeners," *Journal of Medical Speech-Language Pathology*, vol. 7, no. 2, pp. 91–95, 1999.