Published in final edited form as:
Interspeech. 2023 August ; 2023: 2353–2357. doi:10.21437/interspeech.2023-2100.

Combining Multiple Multimodal Speech Features into an Interpretable Index Score for Capturing Disease Progression in Amyotrophic Lateral Sclerosis

Michael Neumann[1], **Hardik Kothare**[1], **Vikram Ramanarayanan**[1]
[1]Modality.AI, Inc., San Francisco, USA

Abstract

Multiple speech biomarkers have been shown to carry useful information regarding Amyotrophic Lateral Sclerosis (ALS) pathology. We propose a two-step framework to compute optimal linear combinations (indexes) of these biomarkers that are more discriminative and noise-robust than the individual markers, which is important for clinical care and pharmaceutical trial applications. First, we use a hierarchical clustering based method to select representative speech metrics from a dataset comprising 143 people with ALS and 135 age- and sex-matched healthy controls. Second, we analyze three methods of index computation that optimize linear discriminability, Youden Index, and sparsity of logistic regression model weights, respectively, and evaluate their performance with 5-fold cross validation. We find that the proposed indexes are generally more discriminative of bulbar vs non-bulbar onset in ALS than their individual component metrics as well as an equally-weighted baseline.

Keywords

speech biomarkers; composite index; multimodal dialog; remote monitoring; clinical trials; amyotrophic lateral sclerosis

1. Introduction

Speech and oro-facial biomarkers have shown great promise for remote assessment and monitoring of neurological and mental health [1, 2, 3, 4]. Indeed, many studies have computed and demonstrated the efficacy of multiple speech metrics that capture how a given disease impacts multiple domains of speech performance – be it motor, anatomical, cognitive, linguistic or affective [1, 5, 6, 7].

A large body of work has further shown the utility of combining biomarkers into an index or optimally weighted combination for clinical practice and pharmaceutical trial applications [8, 9, 9, 10, 11]. Such composite biomarker indexes can provide better diagnostics, discriminative ability and noise robustness than the individual markers alone [12, 13]. However, there is no work systematically examining and comparing different methods of computing interpretable indexes for speech based biomarkers, to our knowledge.

vikram.ramanarayanan@modality.ai .

In this contribution, we present a comparative analysis of different methods of index computation for remotely collected speech biomarkers for Amyotrophic Lateral Sclerosis (ALS). We propose a novel two-step approach toward computing optimal indexes that can be adapted to other domains and diseases. Because index computation on a large set of features, such as those that are typically generated through current state of the art speech processing pipelines, can be computationally expensive, feature selection is an important first step. For this, we use a hierarchical clustering based method to group collinear features together and then select representative features from each cluster based on receiver operating characteristic (ROC) analyses for multiple classification tasks. The second step is the index computation, for which we investigated three different methods: (a) a stepwise distribution-free approach to maximize the Youden J statistic, presented recently by Aznar et al. [12]; (b) Su and Liu's linear discriminant framework [8], which provides an efficient closed-form expression to maximize the area under the ROC curve (AUC) under multivariate normality assumption; and (c) a logistic regression model, which estimates coefficients by minimizing the negative log-likelihood of the observed class labels given the data.

We focus on linear combinations of features because of their computational efficiency, and because interpretability is crucial in the clinical setting. Linear coefficients allow for a straightforward interpretation of the relative importance of each feature. To our knowledge, this is the first investigation of index scores based on multimodal speech features from a large dataset of video recordings from people with ALS (pALS) and healthy controls. For evaluation, we focus on the binary classification *bulbar onset* vs. *non-bulbar onset* (only pALS). This task is useful to assess the indexes' sensitivity at capturing changes in the bulbar subsystem (that controls speech), which occur early on in individuals with bulbar onset.

2. Data

The Modality service, a cloud-based multimodal dialog system [14] , was used to collect video recordings from participants, who engaged in a structured conversation with Tina, a virtual dialog agent. For more details, the reader is referred to [14, 15].

The dialog protocol elicits different types of speech samples, which are inspired by prior work [16, 17, 18, 19] and also utilized in similar remote monitoring efforts [20]. In this work, we focus on (a) read speech (sentence intelligibility test (SIT), 5-15 words; Bamboo reading passage, 99 words), (b) measure of diadochokinetic (DDK) alternating motion rate (rapidly repeating the syllables /pɑtɑkɑ/), and (c) free speech (a picture description task). After dialog completion, participants were asked to fill out the ALS functional rating scale - revised (ALSFRS-R) [21], the standard clinical scale to capture progression in ALS.

Data from 143 pALS (70 females, mean age (SD): 60.4 (10.2) years, 36 pALS with bulbar symptom onset) and 135 age- and sex-matched[1] healthy controls (71 females, mean age (SD): 59.9 (10.3) years) were collected between 2020-11-03 and 2023-02-08 in collaboration with EverythingALS and the Peter Cohen Foundation[2]. The study protocol

[1] A tolerance of ±3 years was considered a match in age.

was granted exempt status by an external Institutional Review Board[3]. The total number of sessions in the dataset is 5,945. Because this is an ongoing project with continuous enrollment of new participants, there is a large variation in the number of sessions per subject and the time between participants' first and last sessions.

Participant sessions were stratified into three groups based on [22]: controls (**CON**; all sessions from healthy controls, n=3,044), bulbar pre-symptomatic (**PRE**, n=1,162), and bulbar symptomatic (**BUL**, n=1,739). PRE and BUL sessions were grouped based on the ALSFRS-R bulbar subscore (three questions on speech, salivation, and swallowing; score ranges from 0 to 12). 62 out of 143 pALS had normal bulbar function (bulbar subscore = 12) at the beginning of the study; all their sessions *prior* to any decline in bulbar subscore were labeled as PRE. All remaining sessions comprise the BUL group.

3. Methods

3.1. General Experimental Setup

All analysis was performed using Python (v3.7.7) and R (v3.6.1). The following open-source Python libraries were used: Pandas (1.3.5 [23, 24]), Numpy (v1.21.5 [25, 26]), scikit-learn (v1.0.2 [27]), Matplotlib (v3.5.1 [28]), and SciPy (v1.7.3 [29]). The following R packages were used: ROCR (v1.0.7 [30]), pROC (v1.14.0 [31], SLModels (v0.1.2 [32]), and the rpy2 interface (v2.9.4 [4]).

We used 5-fold cross validation for the evaluation of the index scores, using `sklearn`'s *StratifiedGroupKFold*. The class labels for stratification were the three cohorts CON, PRE, and BUL. Participant IDs were used as groups to ensure non-overlapping partitions, i.e., for each participant, all sessions are either in the train or in the test partition.

The feature selection step is difficult in a cross validation setup because the feature set is likely to be different for each fold, which hampers comparability and analysis of results. Therefore, feature selection was done on one randomly picked train partition.[5] Sessions with missing values within the selected feature set were dropped from analysis. Before computing index scores and evaluating performance, the data was scaled with `sklearn`'s *MinMaxScaler* (fitted on the train set and applied to the corresponding test set).

3.2. Multimodal Features and Feature Selection

The dialog platform is equipped with analytics modules that extract features from different modalities. Speech and video metrics are computed automatically in real-time during a session. We use Praat [33] and the Montreal Forced Aligner [34] to extract speech metrics. Facial video metrics are based on facial landmarks generated with MediaPipe Face Mesh [35]. Linguistic features are computed for the picture description task only, using the Python

[2] https://www.everythingals.org/research
[3] https://www.advarra.com/
[4] https://github.com/rpy2/rpy2
[5] We did feature selection also on a different train partition to confirm that the set of selected features is not drastically different between folds.

package spaCy[6]; they are based on automatic transcriptions obtained with AWS transcribe.[7] Table 1 shows the metrics that were considered for this study.

For each of the 5,945 sessions, audiovisual metrics were extracted for each speech task in the protocol. Considering all valid task-metric combinations as individual features results in a very large number of features. To handle multicollinear features and identify a good set of representative features, we applied hierarchical clustering on the Spearman rank-order correlations, similar to the approach in [37]. Ward's method was used for clustering and we plotted a dendrogram for visual inspection of the feature clusters. A distance threshold was chosen manually to select clusters that represent sensible feature groupings in terms of the domain (e.g. frequency or timing related speech features) or the area of the face (e.g. features pertaining to jaw movement). Selecting the threshold for splitting clusters can also be done in a data-driven way [37], but we wanted to ensure that every domain is represented individually.

The final feature set should be versatile with respect to different aspects, like progress monitoring in pALS and early diagnosis (classifying between controls and the PRE group). To select one representative metric per cluster, ROC analyses on the individual features were conducted for the following binary classification tasks: *CON vs. all pALS, CON vs. BUL, CON vs. PRE* (utility for early diagnosis and patient-control stratification), *PRE vs. BUL* (progress monitoring), and *bulbar onset vs non-bulbar onset* (useful for stratification, e.g. in clinical trials). In addition, to assess associations between features and the ALSFRS-R bulbar subscore, we estimated the total information coefficient (TICe) and the maximum information coefficient (MICe), using the MICtools software package [38]. The TICe/MICe framework provides good power (finding statistically significant relationships), while being equitable (assigning similar scores regardless of the relationship type, e.g. linear or exponential) [39].

For every cluster, the metric that yielded the best result in the majority of these tests (highest area under curve (AUC) of the ROC curve and highest MICe) was selected as representative. For this, metrics were considered independent of the task first (e.g. identify HNR as best metric for voice quality), and then the task was selected based on majority vote, but also based on the principle of a minimal set of tasks, to reduce participants' burden[8]. For example, when the average jaw center speed metric extracted from the SIT and DDK tasks both performed similarly, DDK was preferred because it was already in the set of tasks selected for other feature groups. In this manner, the tasks were reduced to DDK, picture description and the reading passage. Another desirable property besides predictive power is a high test-retest reliability of the extracted metrics. This was assessed by computing Pearson correlation between the metrics in participants' subsequent sessions, which were recorded at most 7 days apart from each other. The underlying assumption is that within one week we do not expect changes due to disease progression, thus, changes are attributed

[6] https://spacy.io/

[7] https://aws.amazon.com/transcribe/

[8] We made an exception for the lexico-semantic feature cluster. Here, we chose noun-to-verb ratio over the majority-voted verb rate because of its ability to capture information about both verb and noun use, which has been shown to be useful in neurological disorders like ALS [40].

to measurement variability (and variability in the performance of the tasks). The right-hand side of table 1 shows the final feature set along with each feature's test-retest reliability.

3.3. Index Score Computation

We compared three methods of index computation to optimize linear discriminability between cohorts. For evaluation, we focused on the binary classification task to discriminate pALS with bulbar onset vs. those with non-bulbar onset.

As a *baseline* composite index, the features were additively combined with equal weights. However, this simple approach ignores the fact that some features are on average larger in one class, while some are smaller. To illustrate this with an example: PPT is on average higher in pALS than in controls (more and longer pauses), while speaking rate declines and is on average lower in pALS. To account for this, features were inverted by taking (1 – scaled feature) when their mean was smaller in the bulbar onset cohort than in the non-bulbar onset cohort. This was the case for CTA and average mouth symmetry ratio.

3.3.1. Youden Index Optimization—Besides maximizing AUC, another optimization criterion that is commonly utilized is Youden's $\mathcal{J}$ statistic, or Youden index [41]. It is defined as $\mathcal{J} = sensitivity + specificity - 1$. The objective is to find the cut-off point for a diagnostic test or biomarker that maximizes $\mathcal{J}$. Aznar et al. proposed a stepwise distribution-free approach to find the optimal linear combination of continuous biomarkers based on maximizing the Youden index [12]. We used their R package `SLModels` to calculate the feature weights. The advantage of this stepwise approach is that it is non-parametric and distribution-free.

3.3.2. Fisher's Linear Discriminant Function—The method proposed by Su and Liu [8] provides a closed form solution to find the best linear combination that maximizes the AUC, which is based on Fisher's linear discriminant analysis (LDA). Under assumption of Gaussian distributions, the coefficients are proportional to

$$\left(\frac{S_x}{m-1}+\frac{S_y}{n-1}\right)^{-1}(\bar{Y}-\bar{X}) \tag{1}$$

where S_x and S_y are sample covariance matrices of the two cohorts, m and n the respective number of samples in each group, and $\bar{X}$ and $\bar{Y}$ the sample means. This method assumes normal distribution of features, which might not always be the case. However, the calculated index can nevertheless serve as a useful marker, and the method comes with the advantage of low computational cost.

3.3.3. Logistic Regression—We used `sklearn`'s *LogisticRegression* to calculate the model coefficients that serve as feature weights. We used the *liblinear* solver and L1 regularization, and did a grid search cross validation (within each train partition) over the parameter C to optimize for AUC. L1 regularization was chosen because it enforces a sparse weight vector, which is beneficial because a minimal number of features is desired to improve clinical utility [42].

4. Results

Table 2 shows the mean results for the index scores and individual features across 5 folds, including the Youden index and AUC on the train set, and sensitivity, specificity, and UAR on the test set. To obtain results on the test set, we computed the optimal cut-off point that maximizes the Youden index on the train set (using the R package pROC) and applied it as a threshold to classify the test samples.

Speaking duration of the reading passage is the best single feature, which establishes a strong baseline to beat. The Youden $\mathcal{J}$ based method yields a slightly higher test result than speaking duration and all three index scores yield better results than the baseline index (in terms of train AUC and test UAR).

In general, we observed high variance between the individual cross validation folds. For individual features the standard deviation of UAR across folds ranges between 0.036 and 0.096, depending on the feature (0.046 for speaking duration (RP)). For the index scores, the standard deviation was 0.035, 0.037, and 0.048 for logistic regression, LDA, and Youden $\mathcal{J}$ respectively. This suggests that overall the variance is reduced when applying an index as compared to individual features, which supports the use of such an optimal composite index as a relatively more noise-robust composite biomarker.

One of the benefits of linear models for index score composition is their interpretability. It is straightforward to analyze the weights and identify features that contribute the most information to the composite index, an important consideration during deployment in clinic or in drug trials. Figure 1 presents the weights computed with the Youden $\mathcal{J}$ based and the LDA based method. Speaking duration is picked as the most relevant feature in the Youden $\mathcal{J}$ based index and the weights are relatively stable across test folds compared to the LDA based method. In contrast, the LDA based method assigns the largest weights to CTA, and the weights are more unstable across folds. For logistic regression, L1 regularization was used (enforcing coefficient sparsity), setting most feature weights (close) to zero. Features that had non-zero weights across most test folds were CTA, max. lip width, and max. eyebrow displacement.

5. Discussion

In this study we presented a two-step method for selecting relevant features and then combining them into a weighted index score for ALS onset prediction and progress monitoring, which has the potential to improve the utility for clinical practice and pharmaceutical trial applications as compared to multiple individual features. Our findings show that the three investigated methods for index computation – logistic regression, LDA based, and Youden index optimization – all yield indexes that perform better than the baseline of equally weighted features and most of the individual features, while returning lower performance variability overall (and therefore better noise-robustness) across test partitions. Overall, the Youden index based composite score yielded the best result (approximately on par with speaking duration as individual feature). The proposed methods for linear combinations of features are clinically interpretable because the relative contributions of individual features to the overall index are known. Furthermore, the metrics themselves are chosen to be clinically meaningful and interpretable, as opposed to learnt representations.

We analyzed and compared the feature weights and found that the Youden index based and the logistic regression coefficients were more stable across different train/test splits of the data than the LDA based weights, in terms of the relative contribution of each feature. One peculiar difference between the methods was the relative importance of the features speaking duration and CTA. Both features are related (a longer speaking duration leads to lower timing alignment), which might be the reason why CTA was basically disregarded in the Youden index method. In future work, an ablation study, where features are taken out one at a time, can provide more insights about the individual contributions.

As mentioned earlier, one advantage of the LDA based method is the low computational cost. However, it can only provide the optimal linear combination for maximizing the AUC when the features are normally distributed. Despite this not being the case for all features in the dataset we used, we showed that the resulting index score can yield competitive results. Future work will examine other probabilistic variants of this method that do not assume linear separability of classes.

Note that we chose not to provide a single weight vector for the selected feature set and classification task, but instead reported weights for each cross validation fold. An alternative to cross validation is to randomly select one train and one test set. However, this bears the risk of an over-optimistic performance estimate (which becomes clear when looking at the variation between results of cross validation folds). Future work will therefore focus on applying and extending the proposed framework to higher sample sizes (in ALS as well as other disease conditions), which will allow one to produce more generalizable, noise-robust and discriminative indexes, much sought-after as a potential biomarker in clinical care and pharmaceutical trials.

Acknowledgements

This work was funded by the National Institutes of Health grant R42DC019877. We thank all study participants for their time and we gratefully acknowledge the contribution of the Peter Cohen Foundation and EverythingALS towards participant recruitment and data collection.

7. References

[1]. Ramanarayanan V, Lammert AC et al. , “Speech as a Biomarker: Opportunities, Interpretability, and Challenges,” Perspectives of the ASHA Special Interest Groups, vol. 7, no. 1, pp. 276–283, 2022.

[2]. Robin J, Harrison JE et al. , “Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations,” Digital Biomarkers, vol. 4, no. 3, pp. 99–108, 2020. [PubMed: 33251474]

[3]. Milling M, Pokorny FB et al. , “Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell,” Frontiers in Digital Health, vol. 4, 2022.

[4]. Low DM, Bentley KH, and Ghosh SS, “Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review,” Laryngoscope Investigative Otolaryngology, vol. 5, no. 1, pp. 96–116, 2020. [PubMed: 32128436]

[5]. Boschi V, Catricala E et al. , “Connected Speech in Neurodegenerative Language Disorders: A Review,” Frontiers in Psychology, vol. 8, p. 269, 2017. [PubMed: 28321196]

[6]. Rowe HP, Gutz SE et al. , “Acoustic-Based Articulatory Phenotypes of Amyotrophic Lateral Sclerosis and Parkinson’s Disease: Towards an Interpretable, Hypothesis-Driven Framework of Motor Control,” in Proc. Interspeech, 2020, pp. 4816–4820.

[7]. Rowe HP, Shellikeri S et al. , “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.

[8]. Su JQ and Liu JS, “Linear Combinations of Multiple Diagnostic Markers,” Journal of the American Statistical Association, vol. 88, no. 424, pp. 1350–1355, 1993.

[9]. Yu W and Park T, “Two Simple Algorithms on Linear Combination of Multiple Biomarkers to Maximize Partial Area Under the ROC Curve,” Computational Statistics & Data Analysis, vol. 88, pp. 15–27, 2015.

[10]. Pepe MS and Thompson ML, “Combining Diagnostic Test Results to Increase Accuracy,” Biostatistics, vol. 1, no. 2, pp. 123–140, 2000. [PubMed: 12933515]

[11]. Kang L, Liu A, and Tian L, “Linear Combination Methods to Improve Diagnostic/Prognostic Accuracy on Future Observations,” Statistical Methods in Medical Research, vol. 25, no. 4, pp. 1359–1380, 2016. [PubMed: 23592714]

[12]. Aznar-Gimeno R, Esteban LM et al. , “A Stepwise Algorithm for Linearly Combining Biomarkers under Youden Index Maximization,” Mathematics, vol. 10, no. 8, p. 1221, 2022.

[13]. Bansal A and Sullivan Pepe M, “When Does Combining Markers Improve Classification Performance and What Are Implications for Practice?” Statistics in Medicine, vol. 32, no. 11, pp. 1877–1892, 2013. [PubMed: 23348801]

[14]. Suendermann-Oeft D, Robinson A 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.

[15]. Neumann M, Roesler O et al., “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. ISCA, 2021, pp. 4783–4787. [Online]. Available: 10.21437/Interspeech.2021-1801

[16]. Silbergleit AK, Johnson AF, and Jacobson BH, “Acoustic Analysis of Voice in Individuals With Amyotrophic Lateral Sclerosis and Perceptually Normal Vocal Quality,” Journal of Voice, vol. 11, no. 2, pp. 222–231, 1997. [PubMed: 9181546]

[17]. Tomik B and Guiloff RJ, “Dysarthria in Amyotrophic Lateral Sclerosis: A Review,” Amyotrophic Lateral Sclerosis, vol. 11, no. 1-2, pp. 4–15, 2010. [PubMed: 20184513]

[18]. Novotny M, Melechovsky J et al. , “Comparison of Automated Acoustic Methods for Oral Diadochokinesis Assessment in Amyotrophic Lateral Sclerosis,” Journal of Speech, Language, and Hearing Research, vol. 63, no. 10, pp. 3453–3460, 2020.

[19]. Agurto C, Pietrowicz M et al., “Analyzing Progression of Motor and Speech Impairment in ALS,” in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 6097–6102.

[20]. Baxi EG, Thompson T et al. , “Answer ALS, a Large-Scale Resource for Sporadic and Familial ALS Combining Clinical and Multi-omics Data From Induced Pluripotent Cell Lines, year = 2022,” Nature Neuroscience, pp. 1–12.

[21]. Cedarbaum JM, Stambler N 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. [PubMed: 10540002]

[22]. Allison KM, Yunusova Y et al. , “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. [PubMed: 28355886]

[23]. The pandas development team, pandas-dev/pandas: Pandas, Feb. 2020. [Online]. Available: 10.5281/zenodo.3509134

[24]. McKinney W et al., "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, vol. 445, no. 1. Austin, TX, 2010, pp. 51–56.

[25]. Van Der Walt S, Colbert SC, and Varoquaux G, "The NumPy Array: A Structure for Efficient Numerical Computation," Computing in Science & Engineering, vol. 13, no. 2, pp. 22–30, 2011.

[26]. Oliphant TE, "Python for Scientific Computing," Computing in Science & Engineering, vol. 9, no. 3, pp. 10–20, 2007.

[27]. Pedregosa F, Varoquaux G et al. , "Scikit-learn: Machine Learning in Python," The Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[28]. Hunter JD, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[29]. Virtanen P, Gommers R et al. , "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods, vol. 17, no. 3, pp. 261–272, 2020. [PubMed: 32015543]

[30]. Sing T, Sander O et al. , "ROCR: Visualizing Classifier Performance in R," Bioinformatics, vol. 21, no. 20, p. 7881, 2005. [Online]. Available: http://rocr.bioinf.mpi-sb.mpg.de

[31]. Robin X, Turck N et al. , "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves," BMC Bioinformatics, vol. 12, p. 77, 2011. [PubMed: 21414208]

[32]. Aznar-Gimeno R, Esteban LM et al., SLModels: Stepwise Linear Models for Binary Classification Problems under Youden Index Optimisation, 2022, r package version 0.1.2. [Online]. Available: https://CRAN.R-project.org/package=SLModels

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

[34]. McAuliffe M, Socolof M et al. , "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in Proc. Interspeech, 2017, pp. 498–502.

[35]. Kartynnik Y, Ablavatski A et al. , "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

[36]. Liscombe J, Neumann M et al., "On Timing and Pronunciation Metrics for Intelligibility Assessment in Pathological ALS Speech," in Speech Motor Control Conference (SMC), 2022.

[37]. Ienco D and Meo R, "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.

[38]. Albanese D, Riccadonna S et al. , "A Practical Tool for Maximal Information Coefficient Analysis," GigaScience, vol. 7, no. 4, p. giy032, 2018.

[39]. Reshef YA, Reshef DN et al. , "Measuring Dependence Powerfully and Equitably," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 7406–7468, 2016.

[40]. Aiello EN, Pain D et al. , "Rethinking Motor Region Role in Verb Processing: Insights From a Neurolinguistic Study of Noun-Verb Dissociation in Amyotrophic Lateral Sclerosis," Journal of Neurolinguistics, vol. 66, p. 101124, 2023.

[41]. Youden WJ, "Index for Rating Diagnostic Tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950. [PubMed: 15405679]

[42]. Kueffner R, Zach N et al. , "Stratification of Amyotrophic Lateral Sclerosis Patients: A Crowdsourcing Approach," Scientific Reports, vol. 9, no. 1, p. 690, 2019. [PubMed: 30679616]

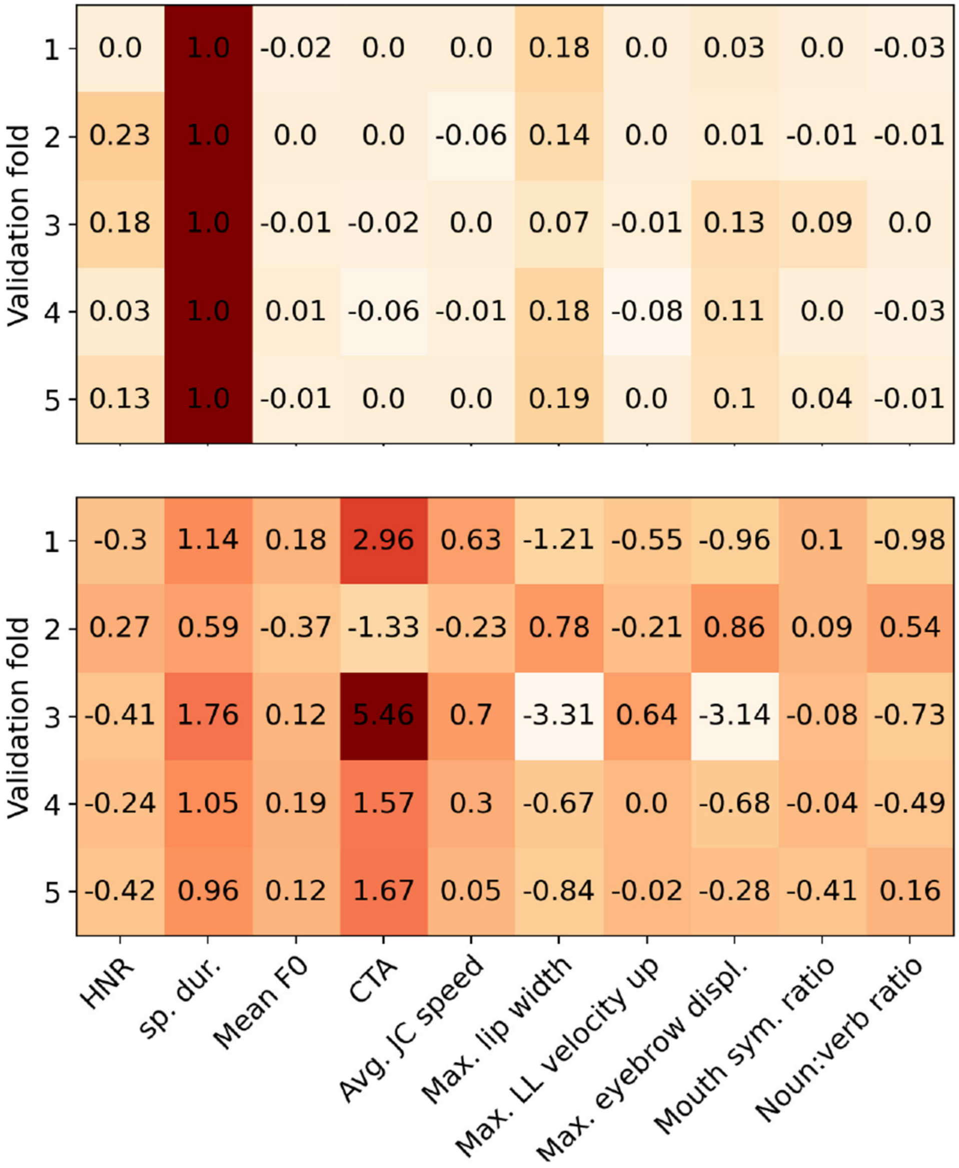

Figure 1:

Feature weights across 5 validation folds for Youden $\mathcal{J}$ based index (top) and LDA based index (bottom).

Table 1:

Overview of extracted metrics, feature clusters and selected representative features. For visual metrics, functionals (minimum, maximum, average) are applied to produce one value across all video frames of an utterance. Visual distance metrics are measured in pixels and are normalized by dividing them by the intercanthal distance (distance between inner corners of the eyes) for each subject. *the text metric word count for the picture description task was clustered together with timing related audio features. TRR: test-retest reliability, LL: lower lip, JC: jaw center, RP: reading passage, PD: picture description, DDK: diadochokinesis

	Metrics	Feature cluster	Selected representative	TRR
Audio	shimmer (%), harmonics-to-noise ratio (HNR, dB), jitter (%)	Voice quality	HNR (DDK)	0.81
	speaking and articulation duration (sec.)*, articulation and speaking rate (WPM)	Duration & Rate	speaking duration (RP)	0.95
	mean, max., min., stdev. of fundamental frequency F0 (Hz)	F0-related	mean F0 (RP)	0.95
	percent pause time (PPT, %), canonical timing alignment (CTA, %) [36], number of syllables (specific for DDK), word count*	Timing alignment	CTA (RP)	0.92
Video	velocity, acceleration, jerk, and speed of jaw center	Jaw movement	avg. JC speed (DDK)	0.73
	lip aperture/opening, lip width, mouth surface area	Mouth measurements	max. lip width (RP)	0.80
	velocity, acceleration, jerk, and speed of lower lip	Lip movement	max. LL velocity upwards (RP)	0.61
	eye opening, vertical displacement of eyebrows	Eyes-related	max. vertical eyebrow displacement (RP)	0.77
	mean symmetry ratio between left and right half of the mouth	Mouth symmetry	avg. mouth sym. ratio (PD)	0.69
Text	percentage of content words, noun rate, verb rate, pronoun rate, noun-to-verb ratio, noun-to-pronoun ratio, closed class word ratio, idea density	Lexico-semantic	verb-to-noun ratio (PD)	0.25

Table 2:

Mean results from 5-fold cross validation for the binary classification task bulbar onset vs. non-bulbar onset. [1] DDK, [2] Reading passage, [3] Picture description. UAR: unweighted average recall, Sen.: sensitivity, Spec.: specificity.

	TRAIN		**TEST**		
	$\mathcal{J}$	**AUC**	**Sen.**	**Spec.**	**UAR**
HNR[1]	0.33	0.69	0.71	0.62	0.67
sp. dur[2]	**0.70**	**0.86**	**0.87**	0.73	**0.80**
mean F0[2]	0.25	0.64	0.43	0.66	0.55
CTA[2]	0.64	0.84	0.14	**0.93**	0.53
avg. JC speed[1]	0.14	0.55	0.71	0.39	0.55
max. lip width[2]	0.30	0.70	0.79	0.48	0.63
max. LL vel. up[2]	0.14	0.58	0.62	0.45	0.54
max. eyebrow displ.[2]	0.34	0.70	0.57	0.73	0.65
avg. mouth sym.[3]	0.06	0.53	0.48	0.43	0.46
verb:noun ratio[3]	0.16	0.54	0.40	0.74	0.57
Baseline	0.53	0.82	0.69	**0.74**	0.71
Youden	**0.75**	0.87	**0.88**	**0.74**	**0.81**
LDA	0.67	**0.88**	0.83	0.70	0.77
Log. regr.	0.67	0.87	**0.88**	0.71	0.79