Harnessing Remote Speech Tasks for Early ALS Biomarker Identification

1st Carla Agurto
IBM Research
Yorktown Heights, NY, USA
carla.agurto@ibm.com

2nd Michele Merler
IBM Research
Yorktown Heights, NY, USA
mimerler@us.ibm.com

3rd Esteban G. Roitberg
EverythingALS
Los Altos, US
esteban@everythingals.org

4th Alan Taitz*
SRI International, Menlo Park, CA, US
**This work was performed while working at EverythingALS*
alan.taitz@sri.com

5th Marcos A. Trevisan
Universidad de Buenos Aires, FCEN, Física
CONICET - UBA, Instituto de Física Interdisciplinaria y Aplicada (INFINA)
Buenos Aires, Argentina.
marcos@df.uba.ar

6th Diego E. Shalom
Universidad de Buenos Aires, FCEN, Física
CONICET - UBA, Instituto de Física Interdisciplinaria y Aplicada (INFINA)
Buenos Aires, Argentina.
diego@df.uba.ar

7th Julian Peller
EverythingALS
Los Altos, US
julian@everythingals.org

8th Lyle W. Ostrow
Department of Neurology, Temple University
Philadelphia, PA, USA
Lyle.Ostrow@tuhs.temple.edu

9th Indu Navar
EverythingALS
Los Altos, USA
indu@everythingals.org

10th Ernest Fraenkel
Massachusetts Institute of Technology
Cambridge, USA
fraenkel@mit.edu

11th James Berry
MGH Institute of Health Professions
Boston, USA
jdberry@mgh.harvard.edu

12th Guillermo A. Cecchi
IBM Research
Yorktown Heights, NY, USA
gcecchi@us.ibm.com

13th Raquel Norel
IBM Research
Yorktown Heights, NY, USA
rnorel@us.ibm.com

***Abstract*—Biomarkers are fundamental for improving early diagnosis, monitoring treatment response, and deepening our understanding of disease mechanisms. The lack of effective biomarkers is particularly detrimental in diseases like amyotrophic lateral sclerosis (ALS), where delays in diagnosis can span 12-18 months, significantly affecting conditions marked by rapid disability progression and reduced lifespan. In this work, we analyzed recordings from 291 participants, including 135 people with ALS (pALS), who performed nine different speech tasks during each session, totaling 6,276 sessions. These recordings were processed using OpenSMILE to extract acoustic features, which were input into three classifiers. We aimed to discriminate pALS from controls and identify different stages of ALS (bulbar manifest and bulbar pre-manifest). We achieved an Area Under the Curve (AUC) of up to 66% (with a recall rate of 79%) and up to 90% (with a recall rate of 91%) for discriminating pre-manifest and manifest ALS from controls, respectively. This work represents a significant step toward identifying reliable biomarkers for ALS, offering new insights into early detection and a better understanding of disease progression.**

***Index Terms*—ALS, Amyotrophic Lateral Sclerosis, Digital Health, biomarkers, large data set, speech**

I. Introduction

Amyotrophic lateral sclerosis (ALS) is a complex neurodegenerative condition that primarily targets the upper and lower motor neurons responsible for voluntary muscle movement. The onset of the disease is marked by symptoms such as limb weakness and muscle spasms, which progress to severe muscle wasting, paralysis, and respiratory failure, typically within 2-4 years of diagnosis [1]–[4].

In clinical settings, ALS is classified based on the family history into either familial or sporadic (no history of ALS in the family) types. Classification is also conducted based on the onset of symptoms into limb (primarily affecting the extremities), axial (with the trunk and neck being the initial sites), bulbar (where speech and swallowing are the primary affected functions), or cognitive (where frontotemporal dementia and ALS present concurrently). It is not uncommon for pALS to present with multiple symptoms at the point of diagnosis, with additional symptoms emerging as the disease progresses. For example, although 30% of pALS initially display bulbar symptoms, dysarthria is estimated to manifest in over 80% of pALS [5].

The Functional Rating Scale-revised (ALSFRS-R) scale [6] is a widely adopted instrument for assessing functional status and disease progression in pALS. The scale is designed to provide a snapshot of a pALS's current abilities in various activities of daily living. It consists of 12 questions that evaluate four functional domains: bulbar (speech, salivation, swallowing), fine motor (handwriting, cutting food and handling utensils, dressing and hygiene), gross motor (turning in bed and adjusting bed clothes, walking, climbing stairs), and respiratory (dyspnea, orthopnea, respiratory insufficiency). For each question, pALS are given five alternatives to select from normal function (score=4) to inability to perform or total loss of function (score=0). Therefore, the total score (adding the partial scores of the 12 questions) can range from 0 (most impaired) to 48 (least impaired). Given our focus on analyzing speech degradation, we used the bulbar domain total score to identify two sub-cohorts in pALS. When pALS have a 12/12 score in the bulbar domain (i.e., normal function), we call them bulbar pre-manifest; otherwise, we call them bulbar manifest.

Researchers have recently focused on detecting pALS from controls based on characterizing speech in specific speech tasks. For example, in [7], the authors analyzed the acoustic components, specifically jitter, and shimmer, in audio recordings of pALS performing a sustained vowel speech task. High accuracy is reported in discriminating between pALS (31) and healthy controls (34). In [8], the focus was on analyzing sustained vowel data by characterizing the harmonic structure of the vowels. Although the participants performed only one speech task, multiple recordings were obtained per participant. By incorporating more complex speech tasks such as reading sentences, along with the TSFEL library [9] and Praat software [10] for audio processing, the authors in [11] achieved high accuracy in discriminating pALS from controls. In [12], authors used a machine learning (ML) approach with speech samples from 123 participants reading the following sentence: "Bamboo walls are getting to be very popular" (one sample per participant). Researchers obtained high AUC performance (greater than 91%) by stratifying sex and ALS status (control:36, symptomatic:51, and presymptomatic:36). In a more recent approach, authors in [13] analyzed the readings from the bamboo passage, acquired through the Winterlight's remote assessment system, using a sparse Bayesian logistic regression classifier achieving AUC of 85% for discriminating pALS (122) from controls (22). Furthermore, it was possible to distinguish bulbar pre-manifest (normal function in the bulbar area) from bulbar manifest with AUC=70%.

Results in the literature are encouraging, but further evaluation is needed on using speech for remote pALS monitoring. For example, the authors in [11] achieved good results but the datasets HomeSenseALS and Minsk used in the analysis contain few participants (less than 50 pALS) or lack of controls (e.g. 15% of controls in [13]), and a reduced set of speech tasks; besides some recordings were captured in a controlled environment (e.g., experiments were performed in the lab or use professional microphones). Additionally, many of these previous studies only contained one recording per participant reducing the inherent variability of human's speech which does not allow to create a model that could be generalizable.

In this study, we overcome these limitations by examining the EverythingALS dataset collected from 291 participants. This includes 135 pALS who completed nine speech tasks during sessions conducted at home. These sessions extended up to more than two years for some participants. We extracted acoustic features using the standard OpenSMILE [14] toolbox to discriminate between pALS vs. control, bulbar manifest vs. bulbar pre-manifest, and bulbar pre-manifest vs. control (early detection).

II. Data Collection

A. Participants

Participants were recruited by the EverythingALS organization (an active ALS community detailed on http://www.everythingals.org), through web advertisements in clinicaltrials.gov, and from flyers that were distributed to specific Neurology clinics. All participants signed an informed consent. In this study, individuals from two populations were recruited: a) individuals diagnosed with ALS or probable ALS and b) individuals with no diagnosis of ALS (non-ALS group). For both groups, the inclusion criteria require that participants must be aged 18 or older, be capable of independently operating a smartphone/tablet/device or PC/laptop, and demonstrate proficiency in reading and speaking English. Participants were further screened to exclude those deemed by the investigator as unable to comply with the study procedure. For the non-ALS group, the inclusion criteria was not having an ALS diagnosis. More details about this study can be found in [15] In what follows, the ALS and non-ALS groups are referred to as the pALS and control groups, respectively. Data used in this analysis was acquired from November 3, 2020, to May 31st, 2023.

B. Protocol

Data from all the participants were collected using the modality.ai platform, a tool designed for the remote evaluation and continuous monitoring of participants. This allows the involvement of participants in weekly verbal exercises, including sustained vowel sounds, reading assignments, diadochokinetic (DDK) rate evaluations, and expressive speech tasks through picture description (PD). To minimize learning biases, the content for the speech tasks, including both sentences and images, was randomly chosen from a pool containing at least 15 sentences per category and 23 unique images. Participants were encouraged to complete the ALSFRS-R scoring form bi-monthly after each session.

III. Audio Data Analysis

A. Feature Extraction

Speech recordings were processed independently by each task to extract acoustic components using the OpenSMILE open source toolbox [14]. We used the Geneva Minimalistic Acoustic Parameter Set (GeMAPS v2.0) [16], consisting of 80 features derived from 18 functionals. This set provides basic acoustic information reflecting physiological changes in voice production. In addition to these features, we have included the duration of the nine tasks in each session.

TABLE I

Demographic and Clinical Characteristics of Study Participants. Distribution of sex, age, race, ethnicity, and ALS-specific clinical measures among all participants, further categorized into the pALS and control groups. Values in all the tables are expressed as median (Q1, Q3).

Category	Variable	All	ALS	Control
Demographics	Participants	291 (100%)	135 (46%)	156 (54%)
	Age (years) at baseline	64 (54, 70)	65 (57, 71)	62 (53, 68)
	Education (in years)	17 (16, 18)	17 (16, 18)	17 (16, 18)
	Sex: Female	173 (59%)	67 (50%)	106 (68%)
	Male	118 (41%)	68 (50%)	50 (32%)
	Race: Caucasian	264 (91%)	123 (91%)	141 (90%)
	African American	4 (1%)	2 (1%)	2 (1%)
	American Indian	6 (2%)	3 (2%)	3 (2%)
	Asian	13 (5%)	5 (4%)	8 (5%)
	Non Reported	4 (1%)	2 (1%)	2 (1%)
	Ethnicity: Not Hispanic	272 (93%)	127 (94%)	145 (93%)
	Hispanic	14 (5%)	5 (4%)	9 (6%)
	Non Reported	5 (2%)	3 (2%)	2 (1%)
	First Language: English	276 (95%)	129 (96%)	147 (94%)
	Other	15 (5%)	6 (4%)	9 (6%)
ALS-specific metrics	Age (years) at symptom onset		61 (53, 67)	
	Months between symptom onset and diagnosis		12 (8, 21)	
	Months between symptom onset and study enrollment		34 (18, 56)	
	Familial/Sporadic type		47 (35%) / 88 (65%)	
	Bulbar/Non-bulbar onset		31 (25.2%) / 101 (74.8%)	
	Condition: Pre-Manifest		74	
	Manifest		41	
	Pre- to Manifest[1]		20	
ALSFRS-R assessment	Total score at baseline		36.3 (7.2)	
	slope[2]		-0.40 (0.48)	
	Bulbar score at baseline		10.4 (1.8)	
	slope[2]		-0.08 (0.23)	
	Speech score at baseline		3.4 (0.8)	
	slope[2]		-0.03 (0.08)	

[1] Number of pALS that were initially Pre-manifest (bulbar score=12) and transitioned to Manifest (bulbar score $<$ 12) during the study.

[2] Only participants with more than three months of data were included.

B. Experimental Design

To evaluate the ability of these features to detect speech deficiencies within pALS, the cohort was split into two groups based on the sum of the three ALSFRS-R bulbar sub-scores (speech, salivation, and swallowing): the pre-manifest group, consisting of individuals with the maximum score (12), for which there is no *self-perceived* speech deterioration, and the pALS-manifest group, comprising individuals with signs of speech impairment, evidenced by ALSFRS-R bulbar scores below 12. Based on these ALS sub-cohorts, we define three classification experiments.

First, we explored whether the acoustic features allow discriminating pALS manifest (pALS with speech deterioration) from controls. Second, we aimed to detect the progression of the disease by distinguishing between pALS manifest and pALS pre-manifest. Finally, the most challenging task was to test the potential of the acoustic features to detect early speech deterioration by segregating pALS pre-manifest from control participants. Taking advantage of the multiplicity of sessions per participant within the EverythingALS dataset, we approached the classification tasks in two ways: (1) using a feature vector per task and (2) using a feature vector per participant, thus averaging over all the participant's sessions.

C. Classification experiments

We tested three different classification algorithms on top of the acoustic features. 1) Extreme gradient boosting (XGboost) [17], one of the most popular classification algorithms for structured data, which implements a scalable distributed gradient-boosted decision tree specifically designed to minimize bias and underfitting. We used the following standard model parameters: $learning\_rate = 0.3$, $\gamma = 0$, $max\_depth = 6$ and $min\_child\_weight = 1$. 2) Light gradient boost machine (LGBM) [18], a gradient boosting tree algorithm similar to XGBoost, growing in a leaf-wise fashion instead of a tree level-wise way, thus increasing training efficiency. We used the following standard model parameters: $learning\_rate = 0.1$, $num\_leaves = 31$, $max\_depth = -1$ (unlimited), $n\_estimators = 100$ and $min\_child\_weight = 1e{-}3$. 3) Support vector machine (SVM) with χ^2 kernel [19], which is a classifier designed to maximize the margin between classes using representative samples (support vectors) in the training data. SVMs can be used with various kernels (linear, RBF, etc.). Still, we adopted the χ^2 one since it has proven very effective for classification experiments, especially with features vectors of reduced dimensionality [19]. We used the following standard model parameters: $C = 1$ and $\gamma = 1/[num_f\, var(X)]$, where num_f is the number of features and $var(X)$ is their variance over the training set.

Features were normalized using min-max normalization before being passed to each classifier, and additional L2 normalization was used for χ^2 SVM. For every classification experiment and session task, the data was split using 10-fold cross-validation to avoid bias. Furthermore, we explicitly balanced the number of examples in each class within the train and test sets for every fold by down-sampling the overrepresented class, which resulted in an almost perfect balance in the number of individuals. Evaluating results on an unbalanced dataset would result in values that do not reflect the real predictive power of the features on metrics such as accuracy or area under the ROC curve (AUC). During training, the data was also sampled to maintain the balance between classes as much as possible and build classifiers without biases toward any class due to data imbalance. Finally, in addition to evaluating the performance of our models, we also assessed their most relevant features by examining their respective weights.

IV. Results

A. Data Collection

Demographics and enrollment information

Table I presents demographics for all participants in the reported period, totaling 291 participants. Of these, 46% (135) were in the pALS group and 54% (156) were in the non-pALS (control) group. Among the pALS, 47 (35%) had familial ALS.

The sex distribution within the pALS was nearly even, with 68 males and 67 females. The control group had a significantly higher female representation at 68%. This disparity is statistically significant (χ^2 test, p-value=0.002). Age differences were also observed, with the control group's median age at baseline at 62 and the pALS group at 65 years (Mann Whitney U test, p-value=0.007). Although the medians and interquartile range for education years are the same, the difference between the distributions was statistically significant (Mann Whitney U test, p-value=0.02). The differences between groups were not statistically significant for the remaining variables described in Table I. Specifically for pALS, the time between symptom onset and study enrollment has a median of 34 months. In addition, we observed that 20 out of 94 individuals of the pALS cohort that initially were pre-manifest (ALSFRS-R bulbar total score 12) transitioned to manifest (ALSFRS-R bulbar total score less than 12).

Task Duration

Table II shows the median duration per task for all participants and by cohort. Individuals in the pALS group took more time to complete almost all the speech tasks than those in the control group. The exceptions were the DDK and PD tasks, where the median for controls was slightly higher than for pALS. All these differences were statistically significant, being the reading tasks the ones with higher discrimination (lower p-values) between groups and the DDK task the one with the lowest discrimination (Mann-Whitney U test, p-value < 0.001). Then, we evaluated whether task duration was also affected for pre-manifest individuals, given their absence of evident speech deterioration. To this end, we subdivided the pALS group into manifest and pre-manifest categories (Figure 1). We found that the difference in task duration across speech tasks between controls and pALS pre-manifest was considerably less than between controls and pALS manifest, as expected. Notably, controls were *slower* than pALS on DDK and PD tasks. Specifically, in comparing controls and pALS manifest, the difference was significant across all nine tasks using a two-sided test. Conversely, in comparing controls to pALS pre-manifest, significance was achieved in only five tasks (SIL-9, SIL-11, and SIL-15, with p-values < 0.01; and Bamboo passage and PD with p-values < 0.001).

TABLE II
Duration in seconds of each speech task for each cohort

Variable	pALS	Control
SIL 5 words	5.5 (5.0, 6.5)	5.2 (4.8, 5.6)
SIL 7 words	5.9 (5.3, 7.1)	5.5 (5.1, 6.1)
SIL 9 words	6.7 (6.0, 8.2)	6.2 (5.7, 6.8)
SIL 11 words	7.9 (6.9, 9.7)	7.2 (6.6, 7.9)
SIL 13 words	8.6 (7.5, 10.8)	7.7 (7.1, 8.5)
SIL 15 words	8.9 (7.8, 11.3)	8.1 (7.5, 8.8)
DDK	18.8 (14.2, 24.1)	19.5 (15.3, 24.6)
Bamboo passage	38.6 (33.4, 51.5)	34.5 (32.0, 37.4)
Picture Description	63.3 (49.8, 67.2)	64.2 (55.9, 67.3)

B. Classification Results

The best performance models for all the classification tasks are shown in Tables III and IV. Performance is higher for features analyzed per participant than per session, with a maximum difference of 11% for SIL-15 words between pALS and controls. Prediction per participant is computed as the median score over the sessions for each participant. Unsurprisingly, the best performance models are obtained when comparing pALS with advanced stages of the disease and controls (pALS manifest vs. controls, AUC $>$ 89.5%). On the other hand, the worst performance models are associated with discriminating pALS pre-manifest from the control cohort (AUC up to 65.9%). Although the differences between speech tasks in the same model (per session and participant) are not statistically significant, models using features derived from DDK perform slightly better than the rest, followed by the task picture description, with the lowest performance model for SIL-7 words. To complement these results, we selected the top features before an inflection point based on weight for each classification experiment and speech task. For most tasks and experiments, the number of features before an inflection point (by sorted weights) is two or three. Results shown in

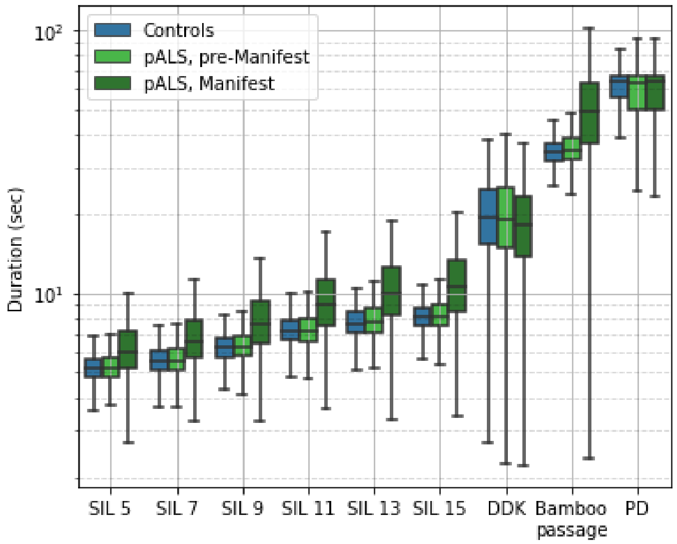


Fig. 1. **Duration of the speech tasks**. The labels of the sentences of increasing length (SIL) are followed by the number of words in each. The other speech tasks consist of the repetition of syllables puh-tuh-kuh (DDK), the reading of the paragraph known as Bamboo Passage (Bamboo), and a picture description (PD). Duration is expressed in a log scale.

Figure 2 indicate that task duration is one of the most frequent features. Another relevant feature in terms of weight in the classifiers and frequency of appearance for the different speech tasks is the Loudness peak rate (called LoudnesSpeakPerSec in the OpenSMILE feature set), especially for the experiments including pALS-manifest as one of the cohorts to discriminate. This feature, a rate of loudness peaks in the participant's speech, can be interpreted as detecting the speech intensity rate. For the experiment pALS pre-manifest vs. controls, the features derived from Mel-frequency cepstral coefficients (MFCCs) are the most relevant. Finally, as the experiment becomes more complex (e.g., pALS pre-manifest vs. control), each speech task's most relevant features are different. For example, when comparing the top 3 features used in the models for all speech tasks (Figure 2), we found there are 23 features for pALS pre-manifest vs. control (a most challenging task). In comparison, there are only seven features for pALS manifest vs. control (a less difficult task), meaning there is good agreement among models.

V. Discussion

We achieved high-performance results in discriminating between pALS manifest and pre-manifest, with AUC up to 90.2%. This outcome is particularly encouraging, considering that 20 subjects in our dataset transitioned from pre-manifest to manifest status, introducing samples into both groups and thereby increasing the complexity of discrimination. We noticed that some tasks, such as DDK, Bamboo passage, and PD, performed better than SIL tasks. A contributing factor could be the duration of the tasks, meaning that longer tasks may allow the extraction of more robust features.

When discriminating controls from all pALS, including individuals with unaffected speech, we achieved AUC values ranging from 78% to 83% at the participant level and slightly lower values (67 % to 74 %) when discriminating based on individual sessions. These results hold up well when compared with those published in previous studies [12], [13], [20]–[22].

Another interesting finding was task duration, a non-direct acoustic property (features extracted in the analysis) of speech. As demonstrated in Figure 1, there is a systematic increase in the duration of the reading tasks for the pALS manifest cohort with respect to the pALS pre-manifest and control cohorts. This suggests that speech becomes slower for the pALS manifest group, which requires more time to complete the reading tasks. A different pattern is observed for the other two types of speech tasks, where controls use more time to finish tasks than pALS. In the DDK task, where participants were instructed to repeat the syllables *puh-tuh-kuh* until running out of breath, this pattern could indirectly reflect diminished lung capacity in pALS. Also, challenges associated with a more difficult task, such as picture description, may contribute to the discrepancy found for this task, with a median duration exceeding one minute. The cognitive load imposed by the PD task must also be considered. Furthermore, although the difference in median duration narrows when comparing pre-manifest pALS to controls, statistical analyses reveal distinct distributions for 5 out of the nine speech tasks. This suggests that task duration could also contribute to identifying early pALS cases. Not surprisingly, when we look at the weight analysis (see Figure 2), we notice that features about task duration and loudness peaks rate (rate of speech intensity peaks) are the most relevant ones, appearing in at least 6 out of 9 speech tasks in the three different classification experiments mentioned above. These features can be seen as proxies for measuring speech rate, which is a key feature for characterizing ALS and has consistently been found in other studies [23]–[26].

The dataset used for analyzing pALS offers unique opportunities for in-depth understanding and innovation, distinguishing it from other datasets. Specifically, the progression rate in our pALS cohort, characterized by a mean decline of ALSFRS-R total score of -0.40, indicates that our cohort predominantly consists of slow progressors. This classification aligns with the progression rate categories defined by Labra et al. [27], with gradients below -1.11 for fast progressors, between -1.11 and -0.47 for intermediate progressors, and slow progressors with gradients exceeding -0.47. Additionally, the mean ALSFRS-R total score of 36.3 at baseline in our pALS cohort suggests a significant representation of individuals at early stages of the disease at the time of recruitment [26], [28]. These characteristics notably increase the challenge of discriminating between pALS pre-manifest and controls in the context of early diagnosis. Particularly, our results for this experiment (pALS pre-manifest and controls) indicate that the most predictive features were related to loudness, shimmer, and formants (F1, F2, F3) across the models. Furthermore, there is less consensus on the relevant features across different

TABLE III

Classification results per session. Performance is evaluated using precision, recall, F1-score, AUC, and Accuracy (mean ± std) for each speech task and classification experiment. Only the classifier with the best performance was displayed and indicated in the last column. The tasks corresponding to reading sentences on increasing length (SIL) are followed by the number of words in each.

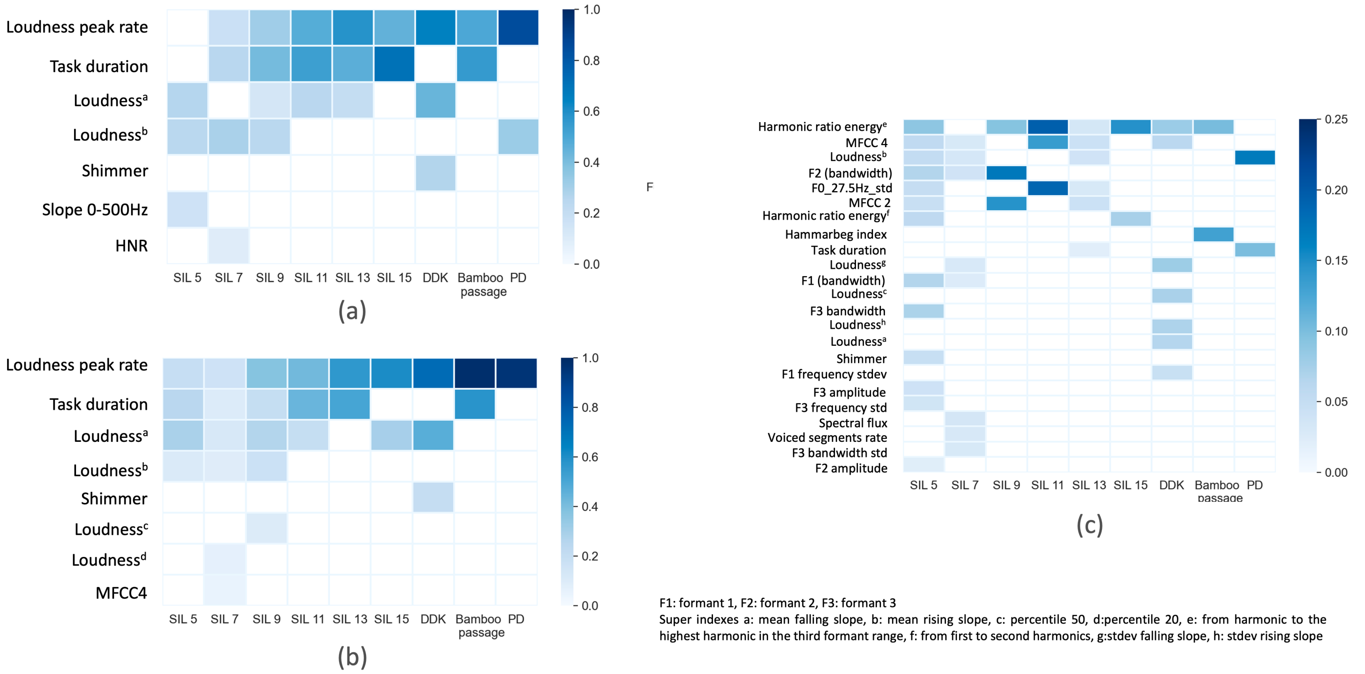
Experiment	Speech Task	Precision	Recall	F1	AUC	Accuracy	Best Classifier
ALS vs. Control	Picture	64.6 ± 6.2	67.8 ± 8.7	65.6 ± 4.3	73.0 ± 5.9	64.7 ± 4.3	SVM
	SIL 5words	63.6 ± 8.6	65.4 ± 8.4	64.4 ± 8.0	69.5 ± 10.4	63.8 ± 8.5	SVM
	SIL 7words	60.3 ± 4.3	64.9 ± 8.4	62.2 ± 4.6	66.7 ± 6.1	60.8 ± 4.3	SVM
	SIL 9words	62.6 ± 8.2	65.5 ± 10.6	63.6 ± 7.7	69.3 ± 9.4	62.7 ± 7.8	SVM
	SIL 11words	64.3 ± 7.5	62.1 ± 11.3	62.7 ± 7.9	70.7 ± 7.3	63.6 ± 6.3	LGBM
	SIL 13words	64.4 ± 6.5	65.3 ± 7.4	64.5 ± 4.9	70.9 ± 8.1	64.1 ± 5.4	SVM
	SIL 15words	65.0 ± 8.4	65.5 ± 14.2	64.6 ± 9.5	71.6 ± 8.7	64.9 ± 7.8	SVM
	Bamboo passage	67.9 ± 12.4	70.9 ± 9.1	68.6 ± 8.1	72.8 ± 11.4	67.0 ± 10.3	SVM
	DDK	**69.4 ± 8.2**	**73.3 ± 6.7**	**70.9 ± 5.4**	**73.9 ± 9.7**	**69.8 ± 6.6**	SVM
Pre-manifest vs. Manifest	Picture	83.1 ± 11.5	77.5 ± 13.5	79.7 ± 11.2	84.9 ± 12.7	77.2 ± 11.1	SVM
	SIL 5words	84.1 ± 12.1	72.8 ± 14.1	77.8 ± 12.9	84.0 ± 10.0	76.5 ± 9.4	LGBM
	SIL 7words	79.5 ± 11.0	73.9 ± 13.0	76.2 ± 10.9	79.9 ± 13.6	73.6 ± 8.8	LGBM
	SIL 9words	87.1 ± 9.4	75.7 ± 14.2	80.5 ± 11.2	85.6 ± 12.0	78.7 ± 10.7	LGBM
	SIL 11words	86.2 ± 9.6	81.2 ± 14.5	83.1 ± 11.2	84.0 ± 16.2	79.5 ± 13.2	LGBM
	SIL 13words	86.4 ± 10.5	79.7 ± 12.5	82.4 ± 10.0	77.9 ± 20.8	79.1 ± 12.1	LGBM
	SIL 15words	85.9 ± 10.9	77.9 ± 10.2	81.3 ± 9.2	87.2 ± 11.7	79.3 ± 8.6	LGBM
	Bamboo passage	**88.9 ± 11.1**	**83.6 ± 12.0**	**85.1 ± 8.0**	86.8 ± 14.9	**81.9 ± 10.1**	LGBM
	DDK	88.3 ± 11.5	81.8 ± 18.7	83.8 ± 13.5	**90.1 ± 11.4**	81.8 ± 14.0	XGBoost
Pre-manifest vs. Control	**Picture**	51.9 ± 5.6	72.4 ± 14.3	60.1 ± 8.0	**57.8 ± 10.2**	52.8 ± 6.8	LGBM
	SIL 5words	51.5 ± 7.2	59.2 ± 13.7	54.7 ± 9.2	54.2 ± 12.5	51.9 ± 8.4	LGBM
	SIL 7words	50.8 ± 8.5	61.1 ± 14.3	55.0 ± 10.2	52.0 ± 13.0	51.0 ± 9.3	LGBM
	SIL 9words	50.0 ± 8.7	61.3 ± 15.7	54.5 ± 11.1	49.9 ± 14.9	50.2 ± 9.6	SVM
	SIL 11words	51.2 ± 8.7	62.0 ± 12.7	55.7 ± 9.4	51.4 ± 14.2	50.9 ± 10.7	SVM
	SIL 13words	50.5 ± 6.1	63.4 ± 8.7	55.8 ± 5.1	50.7 ± 12.8	49.8 ± 7.4	LGBM
	SIL 15words	52.8 ± 5.6	64.6 ± 12.7	57.3 ± 7.0	53.8 ± 12.0	52.8 ± 6.9	XGBoost
	Bamboo passage	51.3 ± 9.3	**73.1 ± 15.9**	60.1 ± 11.4	52.9 ± 17.8	52.1 ± 12.3	XGBoost
	DDK	**53.4 ± 6.3**	71.4 ± 11.6	**60.5 ± 5.8**	54.0 ± 11.6	**53.7 ± 7.2**	XGBoost

TABLE IV

Classification results per subject. Performance is evaluated using precision, recall, F1-score, AUC, and Accuracy (mean ± std) for each speech task and classification experiment. Only the classifier with the best performance was displayed and indicated in the last column. The tasks corresponding to reading sentences on increasing length (SIL) are followed by the number of words in each.

Experiment	Speech Task	Precision	Recall	F1	AUC	Accuracy	Best Classifier
ALS vs. Control	Picture	70.4 ± 9.4	76.4 ± 12.7	72.8 ± 9.3	80.1 ± 9.7	71.8 ± 9.6	SVM
	SIL 5words	71.2 ± 5.4	**83.5 ± 10.3**	76.3 ± 4.8	80.9 ± 7.8	74.4 ± 4.6	SVM
	SIL 7words	66.1 ± 9.3	72.7 ± 10.0	68.7 ± 7.0	78.0 ± 7.1	66.9 ± 8.1	LGBM
	SIL 9words	68.0 ± 8.0	75.9 ± 8.3	71.5 ± 7.2	80.0 ± 7.4	69.7 ± 8.0	SVM
	SIL 11words	66.3 ± 6.4	77.4 ± 12.0	71.1 ± 7.5	79.3 ± 5.5	69.0 ± 7.2	SVM
	SIL 13words	71.4 ± 8.0	73.9 ± 9.0	72.2 ± 6.4	80.6 ± 7.9	71.6 ± 6.9	LGBM
	SIL 15words	74.3 ± 11.3	76.1 ± 14.5	74.1 ± 9.2	**83.2 ± 6.7**	73.9 ± 9.1	SVM
	Bamboo passage	72.8 ± 7.2	76.5 ± 13.2	73.7 ± 7.7	81.4 ± 6.5	73.3 ± 6.4	SVM
	DDK	**75.8 ± 9.4**	81.3 ± 6.0	**78.1 ± 6.0**	81.5 ± 9.8	**76.9 ± 7.3**	SVM
Pre-manifest vs. Manifest	**Picture**	**85.4 ± 12.5**	81.6 ± 13.5	82.5 ± 9.1	88.2 ± 9.6	**83.6 ± 8.7**	SVM
	SIL 5words	80.4 ± 19.8	**91.0 ± 11.1**	**83.3 ± 12.4**	**90.2 ± 8.5**	82.3 ± 13.5	SVM
	SIL 7words	80.4 ± 16.5	78.0 ± 18.0	77.9 ± 14.2	85.0 ± 12.0	79.7 ± 8.3	XGBoost
	SIL 9words	82.2 ± 12.7	86.2 ± 12.4	83.2 ± 8.8	87.0 ± 11.7	82.9 ± 8.9	LGBM
	SIL 11words	78.5 ± 16.9	88.3 ± 13.4	82.4 ± 13.8	88.0 ± 13.6	83.2 ± 12.0	LGBM
	SIL 13words	78.0 ± 17.6	86.5 ± 9.5	81.0 ± 11.9	84.4 ± 14.0	78.9 ± 13.8	LGBM
	SIL 15words	76.8 ± 17.2	83.9 ± 11.7	78.9 ± 10.6	86.4 ± 8.79	79.0 ± 12.3	XGBoost
	Bamboo passage	81.5 ± 17.8	80.7 ± 16.0	79.2 ± 11.8	86.2 ± 16.2	78.8 ± 12.1	LGBM
	DDK	74.6 ± 17.5	82.3 ± 10.2	76.6 ± 10.1	88.4 ± 11.4	78.4 ± 8.3	XGBoost
Pre-manifest vs. Control	**Picture**	51.9 ± 7.6	**79.1 ± 16.6**	62.4 ± 10.1	**65.9 ± 15.0**	53.2 ± 10.6	LGBM
	SIL 5words	55.8 ± 14.4	68.2 ± 25.3	60.6 ± 18.4	62.3 ± 17.2	**58.6 ± 15.1**	SVM
	SIL 7words	53.0 ± 14.0	71.7 ± 19.5	60.6 ± 15.3	55.6 ± 20.9	53.6 ± 17.8	LGBM
	SIL 9words	52.9 ± 18.1	68.0 ± 31.2	58.1 ± 21.8	54.1 ± 14.7	57.7 ± 12.1	XGBoost
	SIL 11words	52.1 ± 10.6	66.4 ± 17.4	57.7 ± 12.3	52.1 ± 17.9	52.0 ± 14.5	SVM
	SIL 13words	56.8 ± 6.95	76.0 ± 11.8	**64.1 ± 4.6**	60.2 ± 12.5	57.8 ± 6.41	LGBM
	SIL 15words	**58.8 ± 11.8**	71.1 ± 19.7	63.2 ± 13.7	63.8 ± 15.2	59.8 ± 14.0	LGBM
	Bamboo passage	54.2 ± 13.4	**79.1 ± 18.0**	63.6 ± 13.9	57.3 ± 20.0	54.8 ± 16.7	LGBM
	DDK	55.9 ± 10.2	76.2 ± 15.9	63.6 ± 9.1	58.4 ± 16.1	56.8 ± 10.6	LGBM

speech tasks, highlighting the complexity involved in this classification experiment. According to the results summarized in



F1: formant 1, F2: formant 2, F3: formant 3
Super indexes a: mean falling slope, b: mean rising slope, c: percentile 50, d:percentile 20, e: from harmonic to the highest harmonic in the third formant range, f: from first to second harmonics, g:stdev falling slope, h: stdev rising slope

Fig. 2. **Frequent Features.** Proportion of folds that have the top features. The top features are the ones before the first inflection point based on the weight of the features for each speech task model using the SVM classifier. a)pALS manifest vs. control, b) pALS manifest vs pALS pre-manifest, and c)pALS pre-manifest vs. control.

Table IV, the highest performance achieved for this experiment had an AUC of 65.9% with a recall of 79.1% for the PD task when classification was performed per subject. Although these results did not reach statistical significance, they suggest that tasks like PD, which assess both motor aspects of speech and cognitive functions, could be particularly suitable in detecting early manifestations in pALS individuals.

Finally, while our dataset includes a considerably higher number of participants and speech samples than other studies, we acknowledge the need for further validation. Future efforts will broaden to analysis by sex and progression rate (i.e., slow, intermediate, and fast) due to the potential variability in speech degradation across these sub-cohorts, as suggested by previous research [12]. Concurrently, we are advancing in extracting features to overcome the limitations of pre-defined feature sets from software packages like the one used in this study. These initiatives aim to refine our understanding and detection of speech deterioration in pALS, contributing valuable insights into the complex nature of ALS progression and its diagnosis.

Acknowledgment

We are profoundly grateful to the study participants for generously donating their data; their invaluable contribution made this work possible.

References

[1] J. Phukan, M. Elamin, P. Bede, N. Jordan, L. Gallagher, S. Byrne, C. Lynch, N. Pender, and O. Hardiman, "The syndrome of cognitive impairment in amyotrophic lateral sclerosis: a population-based study," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 83, no. 1, pp. 102–108, 2012.

[2] L. H. Goldstein and S. Abrahams, "Changes in cognition and behaviour in amyotrophic lateral sclerosis: nature of impairment and implications for assessment," *The Lancet Neurology*, vol. 12, no. 4, pp. 368–380, 2013.

[3] A. Chiò, C. Moglia, A. Canosa, U. Manera, R. Vasta, M. Brunetti, M. Barberis, L. Corrado, S. D'Alfonso, E. Bersano *et al.*, "Cognitive impairment across als clinical stages in a population-based cohort," *Neurology*, vol. 93, no. 10, pp. e984–e994, 2019.

[4] N. Pender, M. Pinto-Grau, and O. Hardiman, "Cognitive and behavioural impairment in amyotrophic lateral sclerosis," *Current Opinion in Neurology*, vol. 33, no. 5, pp. 649–654, 2020.

[5] B. Tomik and R. J. Guiloff, "Dysarthria in amyotrophic lateral sclerosis: A review," *Amyotrophic Lateral Sclerosis*, vol. 11, no. 1-2, pp. 4–15, 2010.

[6] 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.

[7] S. M. Shabber, M. Bansal, and K. Radha, "A review and classification of amyotrophic lateral sclerosis with speech as a biomarker," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–7.

[8] M. Vashkevich and Y. Rushkevich, "Classification of als patients based on acoustic analysis of sustained vowel phonations," *Biomedical Signal Processing and Control*, vol. 65, p. 102350, 2021.

[9] M. Barandas, D. Folgado, L. Fernandes, S. Santos, M. Abreu, P. Bota, H. Liu, T. Schultz, and H. Gamboa, "Tsfel: Time series feature extraction library," *SoftwareX*, vol. 11, p. 100456, 2020.

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

[11] R. Cebola, D. Folgado, A. V. Carreiro, and H. Gamboa, "Speech-based supervised learning towards the diagnosis of amyotrophic lateral sclerosis." in *BIOSIGNALS*, 2023, pp. 74–85.

[12] S. E. Gutz, J. Wang, Y. Yunusova, and J. R. Green, “Early identification of speech changes due to amyotrophic lateral sclerosis using machine classification.” in *Interspeech*, 2019, pp. 604–608.

[13] L. E. Simmatis, J. Robin, M. J. Spilka, and Y. Yunusova, “Detecting bulbar amyotrophic lateral sclerosis (als) using automatic acoustic analysis,” *BioMedical Engineering OnLine*, vol. 23, no. 1, p. 15, 2024.

[14] F. Eyben, M. Wöllmer, and B. W. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor.” in *ACM Multimedia*, A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462. [Online]. Available: http://dblp.uni-trier.de/db/conf/mm/mm2010.html#EybenWS10

[15] I. N. Bingham, R. Norel, E. G. Roitberg, J. Peller, M. A. Trevisan, C. Agurto, D. E. Shalom, F. Aguirre, I. Embon, A. Taitz, D. Harris, A. Wright, K. Seaver, S. Sullivan, J. R. Green, L. W. Ostrow, E. Fraenkel, and J. D. Berry, “Listener effort quantifies clinically meaningful progression of dysarthria in people living with amyotrophic lateral sclerosis,” *medRxiv*, 2024. [Online]. Available: https://www.medrxiv.org/content/early/2024/06/01/2024.05.31.24308140

[16] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[17] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16, 2016, p. 785–794.

[18] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, p. 1.

[19] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06)*, 2006, pp. 13–13.

[20] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, and G. Cecchi, “Detection of amyotrophic lateral sclerosis (als) via acoustic analysis,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September, September 2018.

[21] K. An, M. J. Kim, K. Teplansky, J. R. Green, T. F. Campbell, Y. Yunusova, D. Heitzman, and J. Wang, “Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks.” in *Interspeech*, 2018, pp. 1913–1917.

[22] M. Neumann, O. Roesler, J. Liscombe, H. Kothare, D. Suendermann-Oeft, D. Pautler, I. Navar, A. Anvar, J. Kumm, R. Norel *et al.*, “Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale,” *arXiv preprint arXiv:2104.07310*, 2021.

[23] K. M. Yorkston, “Speech deterioration in amyotrophic lateral sclerosis: Implications for the timing of intervention,” *Jounal of Medical Speech-Language Pathology*, vol. 1, pp. 35–46, 1993.

[24] L. J. Ball, A. Willis, D. R. Beukelman, and G. L. Pattee, “A protocol for identification of early bulbar signs in amyotrophic lateral sclerosis,” *Journal of the neurological sciences*, vol. 191, no. 1-2, pp. 43–53, 2001.

[25] J. Wang, P. V. Kothalkar, M. Kim, A. Bandini, B. Cao, Y. Yunusova, T. F. Campbell, D. Heitzman, and J. R. Green, “Automatic prediction of intelligible speaking rate for individuals with als from speech acoustic and articulatory samples,” *International journal of speech-language pathology*, vol. 20, no. 6, pp. 669–679, 2018.

[26] 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.

[27] J. Labra, P. Menon, K. Byth, S. Morrison, and S. Vucic, “Rate of disease progression: a prognostic biomarker in als,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 87, no. 6, pp. 628–632, 2016. [Online]. Available: https://jnnp.bmj.com/content/87/6/628

[28] F. Kimura, C. Fujimura, S. Ishida, H. Nakajima, D. Furutama, H. Uehara, K. Shinoda, M. Sugino, and T. Hanafusa, “Progression rate of alsfrs-r at time of diagnosis predicts survival time in als,” *Neurology*, vol. 66, no. 2, pp. 265–267, 2006.