Automatic detection of amyotrophic lateral sclerosis (ALS) from video-based analysis of facial movements: speech and non-speech tasks

Andrea Bandini[1], Jordan R. Green[2], Babak Taati[1,3,4], Silvia Orlandi[5], Lorne Zinman[6,7], and Yana Yunusova[1,7,8]

[1] University Health Network: Toronto Rehabilitation Institute, Toronto, Canada
[2] MGH Institute of Health Professions, Boston, USA
[3] Institute of Biomaterials & Biomedical Engineering, University of Toronto, Canada
[4] Department of Computer Science, University of Toronto, Canada
[5] Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada
[6] Neurology, Sunnybrook Healthy Sciences Centre, Toronto, Canada
[7] Brain Sciences, Sunnybrook Research Institute, Toronto, Canada
[8] Department of Speech-Language Pathology, University of Toronto, Canada
andrea.bandini@uhn.ca

***Abstract*—The analysis of facial movements in patients with amyotrophic lateral sclerosis (ALS) can provide important information about early diagnosis and tracking disease progression. However, the use of expensive motion tracking systems has limited the clinical utility of the assessment. In this study, we propose a marker-less video-based approach to discriminate patients with ALS from neurotypical subjects. Facial movements were recorded using a depth sensor (Intel® RealSense™ SR300) during speech and non-speech tasks. A small set of kinematic features of lips was extracted in order to mirror the perceptual evaluation performed by clinicians, considering the following aspects: (1) range of motion, (2) speed of motion, (3) symmetry, and (4) shape. Our results demonstrate that it is possible to distinguish patients with ALS from neurotypical subjects with high overall accuracy (up to 88.9%) during repetitions of sentences, syllables, and labial non-speech movements (e.g., lip spreading). This paper provides strong rationale for the development of automated systems to detect neurological diseases from facial movements. This work has a high social impact, as it opens new possibilities to develop intelligent systems to support clinicians in their diagnosis, introducing novel standards for assessing the oro-facial impairment in ALS, and tracking disease progression remotely from home.**

Keywords: face tracking; Intel RealSense; amyotrophic lateral sclerosis; facial kinematics.

I. Introduction

Amyotrophic lateral sclerosis (ALS) is a fast progressing neurodegenerative disease that affects the motor neurons in the brain, brainstem, and spinal cord and impacts the motor functions of the limbs, trunk, head, and neck muscles [1]. The reported annual incidence of ALS is between 1.5 and 2.5 per 100,000 and prevalence is between 2 to 7 per 100,000 [2]. There are over 30 thousand people living with ALS in North America, and more than 5,600 new cases are diagnosed every year [3]. Although relatively rare, ALS is well known for its devastating consequences – the loss of the ability to ambulate, use hands, speak, swallow, and breathe – and its fast progression. Depending on the site of symptom onset, ALS is classified into bulbar and spinal forms. Bulbar ALS affects speech and swallowing functions at disease onset. Although bulbar functions are affected in only 30% of patients at onset, nearly 80% of those whose symptoms begin purely in the spinal/limb musculature will develop bulbar signs with disease progression [4-6]. The bulbar onset or the emergence of bulbar signs with ALS progression are associated with a short survival (<2 years) [7] and a debilitating disease course. Bulbar symptoms are among the most significant contributors to the reduction of quality of life in ALS. The loss of communication, particularly loss of speech, has been reported as one of the worst aspects related to ALS [8].

In addition to high personal and social costs, ALS is a financially costly disease. Recently, the nation-wide population cost of ALS has been estimated at $1,023 million in the US and individual's annual costs at $64K in the US and $32K in Canada [9]. For bulbar ALS, these costs include assistive technologies for communication, feeding tube placements, and hospitalization costs associated with treating aspiration pneumonia, dehydration, and malnutrition [9].

Detection of bulbar signs, which contributes to the overall disease diagnosis and prognosis, is very challenging. The diagnosis of bulbar ALS is highly subjective and often based on symptom report as well as clinicians' judgement of cranial nerve integrity (including Cranial VII/ Facial nerve) and perceptual evaluation of deviant speech and voice characteristics. However, there is emerging evidence that symptom reports and clinicians' judgments are insensitive to early forms of the disease [10]. The insensitivity of the current assessment methods may be one of the leading reasons for the diagnostic delay in ALS (by ~18 months [11]). An early detection of bulbar signs is highly desirable in order to expedite diagnosis and subsequent access to life-prolonging care, novel therapeutic drug trials, early bulbar management strategies, and augmentative and alternative communication referrals [12]. Moreover, there is a substantial need for novel objective detection methods that are based on accessible technologies and can be easily incorporated into clinical settings.

There is evidence to suggest that the analysis of facial movements during speech tasks may support the detection of ALS [13-16] (Sec. II-A). However, most of the studies conducted so far have used expensive marker-based technology for motion tracking, limiting the investigation only to highly specialized laboratories. Since face movements may be important indicators in the detection and tracking of bulbar ALS, recent developments in low-

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00031

cost depth sensors and computer vision analyses may revolutionize the assessment of this devastating disease.

The aims of this work are to: 1) establish whether the automatic detection of bulbar ALS by using features of facial movements recorded with a 3D depth sensor is feasible; and 2) identify the best speech and non-speech tasks, among the ones commonly used in clinical practice, which are able to discriminate patients with ALS from neurotypical healthy control (HC) subjects. To the best of our knowledge, this is the first study that attempts to detect bulbar ALS using marker-less video-based analysis of facial movements.

II. Related Work

In this section, we provide an overview about the instrumental/clinical examination of facial movements in ALS (Sec. II-A) and video-based approaches for the analysis of facial movements in neurology (Sec. II-B).

A. Instrumental/Clinical assessment of bulbar ALS

Bulbar ALS affects mainly head and neck musculature with negative impact on speech, swallowing, and other oro-motor functions. Although tongue muscles were the most affected in ALS, several studies demonstrated that facial (jaw and lip) movements were sensitive to disease related changes as well [17-20] and were affected earlier than clinical “gold standards” (i.e., speaking rate and speech intelligibility) [14, 15]. Early in the disease, patients with ALS appeared to exhibit increased lip and jaw velocity, compared to HC, which has been attributed to a compensatory response to the tongue movement impairment [10, 21]. However, reduced velocity (slower movements) and longer lip/jaw movement durations prevailed with disease progression [13, 17]. Lower lip and jaw movements showed high sensitivity to bulbar decline indexed by speaking rate and speech intelligibility changes [15, 16]. Furthermore, recent works demonstrate that lip and jaw movements were able to discriminate between early and late stages of bulbar decline in ALS with accuracy up to 86% [22, 23]. These results were obtained by classifying kinematic features of lip and jaw movements (e.g., range, speed, acceleration, jerk, and duration) in a sentence repetition task with a quadratic Support Vector Machine (SVM) classifier. The most sensitive measures of bulbar decline were the path traveled by lower lip, movement duration, speed, acceleration, and jerk peaks of lower lip and jaw [22, 23]. However, the above studies investigated facial movements in ALS using expensive motion capture technology (e.g., optoelectronic techniques, electromagnetic articulography) that limit the clinical utility of this important assessment.

B. Video-based analysis of facial movements in neurology

Many attempts have been made to develop automated video-based approaches for the assessment [24-30] and rehabilitation [31,32] of facial impairments in neurology. Most of the studies to-date have focused on Parkinson’s disease (PD) [24-27] and facial paralysis due to stroke or facial nerve damage [28-30]. In [24], lip movements were investigated by using the Microsoft® Kinect™ and the Supervised Descent Method (SDM) for face alignment [33]. Patients with PD showed reduced velocity and acceleration of lip movements when compared to HC subjects during syllable repetitions. These results were associated with the diagnosis of hypokinetic dysarthria typical of PD. A number of studies [25-27] investigated the role of facial hypokinesia and hypomimia in PD (i.e., the reduced ability to show facial expressions). In [25], facial landmarks were tracked on color videos with SDM and a set of 20 geometric features [34] was extracted at rest (neutral expression) and from acted and imitated expressions. The distance between the feature vector of expressive tasks from the neutral face (index of facial hypomimia) was reduced in patients with PD. The study of facial movements and expressions in PD was also performed in [26, 27] through the detection of Action Units (AUs) [35]. In particular, [27] used geometric and appearance features and SVM to identify the AUs. A reduction of facial expressivity in PD and its relationship with disease severity was demonstrated in [27].

Several approaches were also proposed to assess facial asymmetry due to facial nerve damage and stroke [28-30]. In [28], facial asymmetry was assessed using 3D scans of the face. Asymmetry was measured as the Euclidean distance between corresponding facial points on the original 3D shape and its mirrored version – the higher was the distance, the more extreme was the level of face asymmetry. The method was able to detect facial asymmetry in patients with stroke during the production of a sentence. Facial movements were also assessed for rehabilitation purposes. In [32], the authors proposed an automated Kinect-based approach to provide visual feedback for rehabilitation of facial paralysis. Specifically, a combination of geometric and surface curvature features from 12 facial regions was used to detect the prescribed facial exercises using a random forest classifier.

These findings demonstrated the capabilities of depth sensors and face tracking algorithms for the assessment and treatment of facial impairments in neurological disorders. In this study, we merge the clinical domain knowledge (i.e., the aspects of clinical assessment of the oro-motor/ speech examination) and computer vision analyses of facial movements to develop an automated system for the detection of bulbar ALS.

III. Proposed Method

A. Participants

Ten patients with ALS (6 male, 4 female) and 8 age-matched HC subjects (6 male, 2 female) were recruited for this study. All subjects were native speakers of English and showed normal hearing and no evidence of cognitive impairment on a cognitive screener (score ≥ 26 on the Montreal Cognitive Assessment [36]). Patients were diagnosed with ALS according to the El Escorial Criteria for the World Federation of Neurology [37]. 8 patients reported spinal symptoms, whereas 2 patients presented with bulbar symptoms at onset. The impact of ALS on daily functions was assessed using the ALS Functional Rating Scale – Revised (ALSFRS-R) questionnaire [38]; its bulbar subscore was obtained from speech, swallowing, and salivation symptom report. Clinical and demographic information of the two groups are reported in Tab. I. The study was approved by the Research Ethics Boards at the Sunnybrook Research Institute and UHN: Toronto

Rehabilitation Institute. All participants signed informed consent according to the requirements of the Declaration of Helsinki.

B. Data collection

Subjects were comfortably seated during the recording session. Color and depth videos of the subjects' face were recorded using the Intel® RealSense™ SR300 camera, placed in front of the participants at a distance between 0.4 and 0.5 m from the head. Resolution of both color and depth video streams was 640x480 pixels at 60 frames per second (fps)[1]. Color frames were recorded in 24-bit RGB images (8 bits per channel), whereas depth frames were recorded in 16-bit, one-channel images. Color and depth streams were recorded synchronously, thus the number of frames *N* was the same for color and depth videos recorded during a task. The face was illuminated by a constant and uniform light source placed behind the camera.

During the experiment, each subject was asked to perform a number of speech and non-speech tasks, as reported in Tab. II. These tasks were chosen because they are commonly used by clinicians during the oral motor/ speech examination [39, 40]. At the beginning of the recording session, each subject was recorded at rest (REST) with neutral facial expression for 20 s. The REST task was not considered for the analysis but used for the calculation of baseline parameters used as the reference for some of the geometric features extracted during the speech and non-speech tasks (Sec. II-C). Participants were asked to look at the camera during the recordings, maintaining a frontal position and trying to avoid any excess head movements. Video recordings were performed using the Intel® RealSense™ SDK 2016 R3 and customized C++ code. A separate video recording was captured for each task.

TABLE I. DEMOGRAPHIC AND CLINICAL INFORMATION FOR PARTICIPANTS.

		Age (years)	**Disease duration (months)**	**ALS-FRS (tot. 48)**	**Bulbar subscore (tot. 12)**
ALS	M±SD	61.0±7.5	39.9±26.1	34.8±5.2	9.8±2.1
	RANGE	45-69	18-109	26-40	7-12
HC	M±SD	57.5±9.3	-	-	-
	RANGE	42-68	-	-	-

TABLE II. SPEECH AND NON-SPEECH TASKS RECORDED DURING THE EXPERIMENT.

Type	**Task**	**Description**
Speech Tasks	BBP	Repetition of the sentence "Buy Bobby a puppy" (10 times), at comfortable speaking rate and loudness, briefly pausing between each repetition.
	PA	Repetition of the syllable /pa/ as fast as possible in a single breath
	PATAKA	Repetition of the syllables /pataka/ as fast as possible in a single breath
Non-Speech Tasks	BLOW	Pretend to blow a candle for 5 times
	KISS	Pretend to kiss a baby (5 times)
	OPEN	Maximum mouth opening (5 times)
	SPREAD	Lip spreading (5 times)

C. Pre-processing

For each task, the beginning and end frames of each repetition were identified by visual inspections of the recording performed by a trained observer. The preprocessing steps consisted of (1) face alignment, (2) estimation of 3D facial landmarks, and (3) gap filling and filtering.

1) Face alignment: for each color frame *i* (with $i \in [1,2,\ldots N]$) a vector $\mathbf{p}_i = [x_{i,1},y_{i,1},\ldots.x_{i,49},y_{i,49}]^T$ of 2-dimensional facial points corresponding to eyebrows, eyelids, nose, outer and inner lip contours was detected and tracked using SDM [33]. The pre-trained Matlab implementation of this algorithm published by Xiong & De la Torre [33] was used.

2) 3D facial landmarks: The corresponding *i-th* depth frame was registered to the color image and then sampled in the pixel locations contained in $\mathbf{p}_i$, in order to obtain a vector of 49 scalar values of depth $\mathbf{Z}_i = [Z_{i,1},\ldots Z_{i,49}]^T$. The 3D coordinates of the facial landmarks for the *i-th* frame were estimated using the intrinsic camera parameters, thus obtaining a 3-dimensional vector of points $\mathbf{P}_i = [X_{i,1},Y_{i,1},Z_{i,1}, \ \ldots X_{i,49},Y_{i,49},Z_{i,49}]^T$, where *X* represents the lateral axis, *Y* the vertical axis and *Z* the frontal axis.

3) Gap filling and filtering: The Z value of some facial landmarks (especially mouth corners, located on the border of mouth aperture) could be null due to holes in the reconstructed depth image. In these cases, the trajectories of the points were filled using spline interpolation. Subsequently, the *X*, *Y*, and *Z* trajectories were smoothed using a 5-point moving average window.

D. Feature extraction

The 3D trajectories of the following facial landmarks were considered for the extraction of features: Nose Tip (*NT*), Upper Lip (*UL*), Lower Lip (*LL*), Right and Left mouth Corners (*RC* and *LC*) (Fig. 1). A combination of the features that have already demonstrated their sensitivity to ALS [13-16, 24, 25] and novel measures was considered based on the above landmarks (see summary of the features in Tab. III). Here we chose measures that can be collected across both speech and non-speech tasks in order to perform the between-task comparison. Interpretable decisions are crucial in clinical applications; thus, we

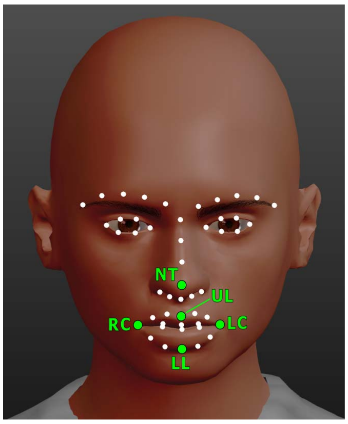


Fig. 1. Facial landmarks obtained with SDM [32] (white circles). The face points of interest are highlighted in green. Face illustration obtained with MakeHuman v.1.1.1 [46].

[1] This is the nominal frame rate as specified by the manufacturer. The actual framerate during the synchronized streaming of color and depth frames ranged between 48 and 51 fps.

avoided automatic feature selection (e.g. using autoencoders) and methods that could process the raw trajectories of these points (e.g. recurrent neural networks). Instead, an engineered feature set was generated based on clinical domain knowledge. Since one of our objectives was the automatic detection of bulbar ALS from video recordings, we used measures that could mirror the clinical evaluation performed by a clinician. Specifically, we considered the following aspects of lip movements: (1) range of motion, (2) speed of motion, (3) symmetry, and (4) shape [39].

1) Features representing range of motion: According to [14], one of the most important facial features that is used to discriminate ALS from HC subjects and detect different stages of ALS bulbar decline is the cumulative path traveled by *LL* during speech movements ($LLpath_{cml}$). This measure is usually higher in patients with ALS, reflecting more laborious articulatory gestures [14]. In this work, $LLpath_{cml}$ was calculated as follows:

$$LLpath_{cml} = \sum_{i=1}^{N} |LLdisp_i - LLdisp_0| \tag{1}$$

where $LLdisp_i$ is the displacement of *LL* at the *i-th* frame of the repetition, and $LLdisp_0$ is the displacement of *LL* at the beginning of the repetition. For each frame of a video-recording, *LLdisp* was calculated as the Euclidean distance between *NT* and *LL*, in order to remove the effect of head rotations and translations from the lip movements. Other range of motion features were the mean value of the mouth area with respect to the rest position (A'_{mean}) and its range ($\Delta A'$) within a repetition. A'_{mean} was defined as:

$$A'_{mean} = \frac{1}{N} \sum_{i=1}^{N} \frac{A_i - A_{rest}}{A_{rest}} \cdot 100 \tag{2}$$

where A_i is the mouth area calculated for the *i-th* frame of a speech or non-speech repetition and A_{rest} is the average mouth area calculated from 10 s of the REST task (the central part from 5 s to 15 s was considered). For each frame, A_i was calculated as the sum of the areas of two triangles defined by *RC*, *UL*, *LL* (right mouth area – A_R) and *LC*, *UL*, *LL* (left mouth area – A_L).

2) Features representing the speed of motion: The first derivative of mouth width (*W* – Euclidean distance between *RC* and *LC*) and *LLdisp* with respect to time was computed to represent the velocity of lip movements on the lateral and vertical axes (*vW* and *vLL*, respectively). Maximum (vW_{max}, vLL_{max}) and minimum velocities (vW_{min}, vLL_{min}) were considered as features for the classification. vW_{max} and vLL_{max} represented the peak velocity of lips during spreading and opening movements, respectively, whereas vW_{min} and vLL_{min} (negative values) were the peak velocity of lips during puckering and closing gestures. As previously demonstrated, patients with ALS tend to have reduced peak velocity of *LL* during opening and closing movements, especially in the later stages of the disease [22, 23].

3) Features representing symmetry: The absolute difference between A_R and A_L was computed for each frame, and its mean value within a repetition ($A_{absdiff}$) was considered as an index of symmetry. Moreover, the Pearson's correlation coefficient between the displacements of *RC* and *LC* (calculated as the Euclidean distance between *RC* and *NT*, and *LC* and *NT*, respectively) was considered as an additional symmetry index of the lower facial muscles (r_{symm}).

4) Features representing the shape/geometry of lips: the roundedness of lips was calculated as the eccentricity (*e*) of an ellipse having mouth width (*W*) and opening (*O* – Euclidean distance between *UL* and *LL*) as axes. For the *i-th* frame of a repetition, the eccentricity was computed as:

$$e_i = \begin{cases} \sqrt{1 - \frac{W_i^2}{O_i^2}}, & W_i < O_i \\ \sqrt{1 - \frac{O_i^2}{W_i^2}}, & W_i > O_i \end{cases} \tag{3}$$

The mean (e_{mean}) and range ($\Delta e = e_{max} - e_{min}$) values within a repetition were used as features for the classification. The value of *e* ranges between 0 and 1, being 0 when $O = W$ (rounded lips), and tends to 1 when $W \gg O$ or $O \gg W$.

E. ALS detection

Each speech and non-speech repetition was represented with an 11-dimensional feature vector. Each feature was standardized using z-scores. A class label *L* was associated with each instance of the dataset: $L = 0$ if a repetition came from HC subjects, $L = 1$ if a repetition was produced by patients with ALS. Two classification algorithms – logistic regression and SVM – were used to evaluate the performance of the proposed feature set during speech and non-speech tasks. Two kernels were used with SVM: linear and radial basis function (RBF). The penalty parameter *C* and the kernel scale parameter γ (for RBF kernel) were chosen through grid search over the possible pairs of *C* and γ (with $C = [2^{-5}, 2^{-3}, \ldots, 2^{15}]$, and $\gamma = [2^{-15}, 2^{-13}, \ldots, 2^{3}]$) [41, 42]. The hyper parameters *C* and γ that produced the lowest classification errors using the tests reported in Sec. IV-B were selected.

TABLE III. SUMMARY OF GEOMETRIC AND KINEMATIC FEATURES EXTRACTED IN THIS STUDY.

Type	Feature	Description
Range of Motion	$LLpath_{cml}$	Cumulative path traveled by *LL* during a repetition
	A'_{mean}	Mean value of mouth area with respect to the rest position
	$\Delta A'$	Range (max-min) of mouth area with respect to the rest position
Speed of Motion	vW_{max}	Maximum velocity of *W* variations
	vW_{min}	Minimum velocity of *W* variations
	vLL_{max}	Maximum velocity of *LL* displacement
	vLL_{min}	Minimum velocity of *LL* displacement
Symmetry	$A_{absdiff}$	Absolute difference between A_R and A_L
	r_{symm}	Correlation between *RC* and *LC* movements
Shape/ Geometry	e_{mean}	Mean value of mouth eccentricity
	Δe	Range (max-min) of mouth eccentricity

IV. EXPERIMENTS AND RESULTS

A. Experimental data

The experiments were performed on a total of 1003 repetitions (634 speech and 369 non-speech): 171 BBP repetitions (91 from patients with ALS, 80 from HC subjects), 210 PATAKA repetitions (98 ALS, 112 HC), 253 PA repetitions (133 ALS, 120 HC), 95 BLOW gestures (51 ALS, 44 HC), 93 KISS gestures (52 ALS, 41 HC), 90 OPEN gestures (50 ALS, 40 HC), and 91 SPREAD gestures (53 ALS, 38 HC). The average number of repetitions per subject was lower in the ALS group, as some of the patients with ALS did not complete the tasks due to fatigue.

B. Classification performance

Classification performance was assessed using a leave-one-subject-out cross validation (LOSO-CV) [43]. The performance was evaluated in terms of *repetition-based* classification, aimed at predicting the label for single repetitions, and *subject-based* classification, where the label of each participant was predicted based on his/her own repetitions.

1) Repetition-based classification: For each iteration of the LOSO-CV, repetitions produced by one participant were considered as test-set and the remaining data used as training set. As a result, every participant (HC and ALS) and his/her repetitions was considered as the test-set. During this test, individuals speech and non-speech repetitions were classified as belonging to the normal or ALS group according to the features reported in Tab. III.

2) Subject-based classification: At each iteration of the LOSO-CV, each subject was considered as test and classified either as normal or ALS, by using the majority vote among the predicted repetitions of the test set. In the case of ties (e.g., number of instances classified as ALS was equal to the number of instances classified as normal), the belonging class was considered as normal, in order to produce a more conservative prediction.

C. Results

The accuracies of prediction (aka the percentage of correctly classified instances relative to the entire number of instances) for each classifier and each task are reported in Fig. 2. The best result in the detection of ALS was obtained in the BBP task using logistic regression during the subject-based classification. Sensitivity (i.e., percentage of ALS instances correctly classified as ALS) and specificity (i.e., percentage of HC instances correctly classified as HC) for this task were 80% and 100%, respectively (Tab. IV). The next best results in terms of classification accuracy were obtained for SPREAD (83.5% during repetition-based classification, 83.3% subject-based classification) and PATAKA (83.3% during subject-based classification), both obtained with linear SVM. The remaining results reported accuracy lower than 80%. Both speech and non-speech tasks reported comparable prediction results, with high sensitivity (90%) during subject prediction.

When the β coefficients of the logistic regressions were examined, the strongest predictors of group membership (highest absolute values) were $LLpath_{cml}$ in BBP ($\beta = 8.82$, $p < .001$) and BLOW ($\beta = 3.03$, $p < .01$), Δe in PATAKA ($\beta = -3.92$, $p < .01$), OPEN ($\beta = -5.98$, $p < .01$), and SPREAD ($\beta = 11.72$, $p < .05$), vLL_{max} in PA ($\beta = 1.72$, $p < .01$), and A'_{mean} in KISS ($\beta = -2.85$, $p < .001$).

The complete list of classification performance (accuracy, sensitivity, and specificity) for each task is reported in Tab. IV. Confusion matrices for the best three tasks - BBP, PATAKA, and SPREAD - are shown in Fig. 3.

D. Discussion

For the first time, our results demonstrate that it was possible to discriminate patients with ALS from HC subjects with high overall accuracy (~90%) by using a small set of geometric and kinematic features of facial (lip) movements extracted with a 3D camera. The best results in terms of accuracy, sensitivity and specificity were obtained with linear models (logistic regression and linear SVM, Fig. 2 and Tab. IV). This might be due to the feature set chosen for this study that allowed a linear separation in the feature space.

Given the relatively limited dataset, we chose a small feature set of interpretable features, encoding geometric and kinematic measures that mirror the assessment performed by clinicians during oro-motor/ speech examination. The expansion of the dataset will allow exploring other facial features not considered here whose sensitivity to ALS and bulbar function decline was already demonstrated (e.g., peaks of acceleration and jerk of lower lip and jaw, which describe the slowdown and smoothness of movements) [22, 23].

The logistic regression coefficients suggested that the most important predictors were features related to range of motion ($LLpath_{cml}$ and A'_{mean}), shape (Δe) and lip velocity (vLL_{max}). The absence of symmetry features is not surprising as orofacial musculature is typically affected bilaterally by ALS. In the future, further investigation of these features will be performed cross-sectionally and longitudinally to analyze the effect of bulbar ALS on orofacial kinematics at different stages of the disease.

TABLE IV. ACCURACY, SPECIFICITY AND SENSITIVITY OF SPEECH AND NON-SPEECH TASKS. FOR EACH TASK, THE RESULTS OBTAINED WITH THE BEST CLASSIFIER ARE REPORTED (^{1}LOGISTIC REGRESSION, ^{2}LINEAR SVM, ^{3}SVM WITH RBF KERNEL).

Task	Validation	Accuracy (%)	Sensitivity (%)	Specificity (%)
BBP^1	Repetition	74.3	78.0	70.0
	Subject	88.9	80.0	100.0
PATAKA^2	Repetition	73.3	79.6	67.9
	Subject	83.3	90.0	75.0
PA^2	Repetition	67.6	69.2	65.8
	Subject	77.8	90.0	62.5
BLOW^2	Repetition	61.1	70.6	50.0
	Subject	66.7	80.0	50.0
KISS^2	Repetition	55.9	100.0	0.0
	Subject	55.6	100.0	0.0
OPEN^3	Repetition	66.7	74.0	57.5
	Subject	72.2	80.0	62.5
SPREAD^2	Repetition	83.5	88.7	76.3
	Subject	83.3	90.0	75.0

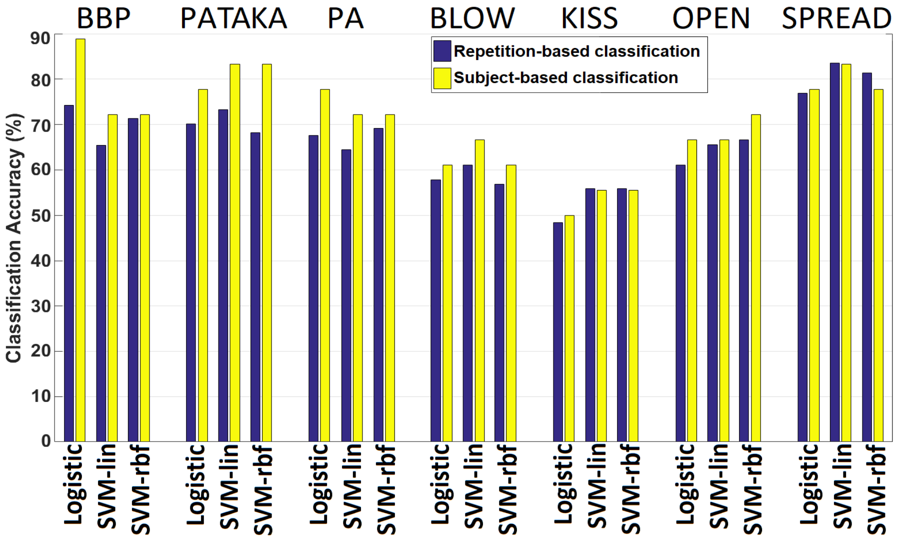


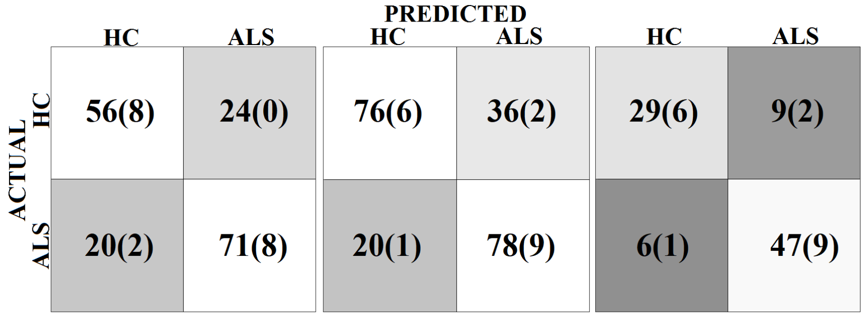
Fig. 2. Classification accuracy during speech and non-speech tasks (Logistic: logistic regression, SVM-lin: linear SVM, SVM-rbf: SVM with RBF kernel).

In this paper, we used SDM as the face alignment step, because of its ability to generalize to untrained situations such as asymmetric face movements [33]. In general, we believe that a non-parametric approach such as SDM is more suitable for studying facial impairments due to neurological diseases, as it allows a higher flexibility in tracking non-typical facial movements. However, using the implementation proposed by Xiong & De La Torre [33], we were only able to focus on lip movements at this time. In the future, additional face alignment algorithms will be tested and compared [44, 45], in order to assess the sensitivity of different approaches to detection of ALS in jaw/face movements.

The sensitivity of kinematic features of lips and jaw to the decline of bulbar functions was demonstrated in the past [13-21]. Recent papers [22, 23] reported accuracy of up to 86% in classifying different stages of bulbar decline using kinematic features of lips and jaw during the BBP task. However, none of the studies conducted so far made use of video-based marker-less methods, which are highly clinically feasible. Thus, our work provides two important contributions: 1) it supports the previous results regarding the importance of facial assessment in the detection of bulbar ALS; and 2) promotes further studies of facial musculature changes in ALS using low-cost marker-less techniques. These results provide strong rationale for the development of affordable technology for assessment and rehabilitation of neurological disorders.

Another innovative contribution of our work is the investigation of several tasks involving facial movements, in order to detect the most predictive movement types in the detection of ALS. The highest classification performance was obtained during the repetitions of sentences and syllables ("speech tasks" BBP and PATAKA) and during the spreading of lips ("non-speech" SPREAD). The best result in terms of prediction of the single repetitions was obtained during the SPREAD task (accuracy 83.5%), whereas the highest accuracy classifying individual subjects was obtained during the repetition of a sentence (BBP, accuracy 88.9%). For most of the tasks, the prediction of the subject class (ALS or normal) using the majority vote among the predicted repetitions of that subject resulted in improved classification performance (Fig. 2). This means that the majority of repetitions produced by individual participants were classified correctly.

At a first inspection, the good results obtained with SPREAD might be surprising, especially considering that this task was rarely used in previous research papers. However, lip spreading may be classified as a labial



	PREDICTED					
ACTUAL	HC	ALS	HC	ALS	HC	ALS
HC	56(8)	24(0)	76(6)	36(2)	29(6)	9(2)
ALS	20(2)	71(8)	20(1)	78(9)	6(1)	47(9)
	a) BBP		b) PATAKA		c) SPREAD	

Fig 3. Confusion matrix for a) BBP task (logistic regression), b) PATAKA task (linear SVM), and c) SPREAD task (linear SVM). The number of repetitions classified during the LOSO-CV is reported outside the brackets, the number of subjects (obtained during the Normal/ALS prediction test) is reported in parentheses.

“maximum performance” task; the lateral lip movements during spread are purely labial and not driven by the jaw that can exert greater forces and often “takes over” the facial muscles in tasks such as mouth opening and speech, particularly in neurological diseases such as ALS. Thus, it might be more likely to find signs of facial muscle weakness in this type of task. On the other hand, speech tasks have clear advantages for the assessment of oro-facial kinematics. For example, video analysis of facial movements may be coupled with the acoustical analysis of speech signals in order to investigate movements of structure (e.g., tongue movements) not visible from video recordings [43]. Since speech and non-speech tasks bring their own advantages (more information during speech tasks, simplicity of gestures during non-speech tasks) to the assessment process, their combination should be considered for future investigations aiming at disease detection.

In this work we did not consider the order of the repetitions and thus the variability of facial movements that may arise in the presence of a neurodegenerative disease. This consideration must be taken into account in future studies, because many patients with ALS did not complete the tasks due to fatigue [15]. It is possible that the latest repetitions may discriminate patients with ALS better than those at the beginning of the task (i.e., the higher the repetition number, the higher the weight that should be given to that repetition).

Since our results revealed higher sensitivity than specificity (see Tab. IV), ALS instances were recognized more readily than HC productions. The number of HC participants was lower than the patients with ALS in our study. This needs to be remediated in future research, which should include a larger normative dataset considering extensive normal variability in speech kinematics. We are currently recruiting more HC subjects as well as patients with ALS. Despite a number of limitations, the overall results of this study are promising as they may lead to the development of a cheap, easy to use, and clinically feasible system that will support clinicians in the assessment and diagnosis of the bulbar form of ALS.

V. Conclusion

In this study we demonstrated that video-based analysis of facial movements is a feasible and accurate way to perform the detection of bulbar ALS. In particular, the most accurate tasks in the prediction of the disease were the repetition of sentences, syllables, and labial movements (lip spread), with accuracies between 80-90%. The detection of bulbar signs of ALS in a clinical setting is still highly subjective and non-standardized. This leads to delayed diagnosis and, therefore, delayed interventions with reduced possibilities to slow down the rapid reduction of quality of life for these patients. This work provides strong rationale to the development of automated marker-less systems for the detection bulbar ALS from facial movements.

The proposed approach can support clinicians in the diagnosis of bulbar ALS, in order to obtain more accurate diagnosis of the disease as a whole as early as possible and introducing novel standards in the assessment of bulbar signs. Patients can also benefit from this automated assessment, as an early diagnosis implies earlier interventions to maintain acceptable standards of quality of life during disease progression. The future implementation of the proposed video-based approach in clinical and home settings will help in reducing the costs related to ALS (over $1 billion in the US only), as an improvement in the detection of bulbar signs may allow the programming of timely interventions to support communication. In this regard, a future challenge will be to test the feasibility of an automated video-based assessment of bulbar ALS at home.

Acknowledgment

The authors would like to acknowledge the funding support from NIH grant R01DC013547, AGE-WELL NCE Inc., and from the University Health Network – Toronto Rehabilitation Institute. We are thankful to the participants and their families for taking part in this project. We also thank Madhura Kulkarni for her assistance with this project.

References

[1] M. C. Kiernan et al., "Amyotrophic lateral sclerosis," *The Lancet*, vol. 377, no. 9769, Mar. 2011, pp. 942-955.

[2] R. H. Brown and A. Al-Chalabi, "Amyotrophic lateral sclerosis," *The New England Journal of Medicine*, vol. 377, no. 2, Jul. 2017

[3] http://www.alsa.org [Accessed: September 29, 2017]

[4] J. T. Caroscio et al. "Amyotrophic lateral sclerosis. Its natural history," *Neurologic Clinics*, vol. 5, no. 1, Feb. 1987, pp. 1-8. pp. 162-172.

[5] A. Chen and C. G. Garrett, "Otolaryngologic presentations of amyotrophic lateralsclerosis," *Otolaryngology-Head and Neck Surgery*, vol. 132, no. 3, Mar. 2005, pp. 500-504.

[6] E. K. Hanson, K. M. Yorkston, and D. Britton, "Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment, and augmentative and alternative communication options," *Journal of Medical Speech-Language Pathology*, vol. 19, no. 3, Sep. 2011, pp. 12-30.

[7] H. Mitsumoto and M. Del Bene, "Improving the quality of life for people with ALS: the challenge ahead," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 1, no. 5, Dec. 2000, pp. 329-336.

[8] M. Hecht et al., "Subjective experience and coping in ALS," *Amyotrophic Lateral Sclerosis and Other Motor Neuron Disorders*, vol. 3, no. 4, Jan. 2002, pp. 225-231.

[9] J. Larkindale et al., "Cost of illness for neuromuscular diseases in the United States," *Muscle & Nerve*, vol. 49, no. 3, Mar. 2014, pp. 431-438.

[10] K. M. Allison 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, Mar. 2017, pp. 1-9.

[11] M. R. Turner et al., "The diagnostic pathway and prognosis in bulbar-onset amyotrophic lateral sclerosis," *Journal of the Neurological Sciences*, vol. 294 no. 1, Jul. 2010, pp. 81-85.

[12] J. R. Green et al., "Bulbar and speech motor assessment in ALS: Challenges and future directions," *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, vol. 14, no. 7-8, Dec. 2013, pp. 494-500.

[13] Y. Yunusova et al., "Articulatory movements during vowels in speakers with dysarthria and healthy controls," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, Jun. 2008, pp. 596-611.

[14] Y. Yunusova et al., "Kinematics of disease progression in bulbar ALS," *Journal of Communication Disorders*, vol. 43, no. 1, Jan-Feb. 2010, pp. 6-20.

[15] P. Rong et al., "Predicting early bulbar decline in amyotrophic lateral sclerosis: A speech subsystem approach," *Behavioural Neurology*, Jun. 2015.

[16] P. Rong et al. "Predicting speech intelligibility decline in amyotrophic lateral sclerosis based on the deterioration of individual speech subsystems," *PloS one*, vol. 11 no. 5, May 2016, e0154971.

[17] S. E., Langmore and M. E. Lehman, "Physiologic deficits in the orofacial system underlying dysarthria in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 37 no. 1, Feb. 1994, pp. 28-37.

[18] A. S. Mefferd, J. R. Green, and G. Pattee, "A novel fixed-target task to determine articulatory speed constraints in persons with amyotrophic lateral sclerosis," *Journal of Communication Disorders*, vol. 45, no.1, Feb. 2012, pp. 35-45.

[19] A. S. Mefferd, G. L. Pattee, and J. R. Green, "Speaking rate effects on articulatory pattern consistency in talkers with mild ALS," *Clinical Linguistics & Phonetics*, vol. 28, no. 11, Nov. 2014, pp. 799-811.

[20] Y. Yunusova, G. G. Weismer, and M. J. Lindstrom, "Classifications of vocalic segments from articulatory kinematics: Healthy controls and speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, Oct. 2011, pp. 1302-1311.

[21] S. Shellikeri et al., "Speech movement measures as markers of bulbardisease in amyotrophic lateral sclerosis," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 5, Oct 2016, pp. 887-899.

[22] A. Bandini et al., "Classification of bulbar ALS from kinematic features of the jaw and lips: Towards computer-mediated assessment," in Proc. *Interspeech 2017*, Stockholm, Sweden: ISCA, 2017, pp. 1819-1823.

[23] A. Bandini et al., "Kinematic features of jaw and lips distinguish symptomatic from pre-symptomatic stages of bulbar decline in amyotrophic lateral sclerosis (ALS)," *Journal of Speech, Language, and Hearing Research*, to be published.

[24] A. Bandini et al., "Markerless analysis of articulatory movements in patients with Parkinson's disease," *Journal of Voice*, vol. 30, no. 6, Nov. 2016, pp 766.e1-766.e11.

[25] A. Bandini et al., "Analysis of facial expressions in Parkinson's disease through video-based automatic methods," *Journal of Neuroscience Methods*, vol. 281, Apr. 2017, pp. 7-20.

[26] N. Vinokurov et al., "Quantifying Hypomimia in Parkinson Patients Using a Depth Camera," in Proc. *International Symposium on Pervasive Computing Paradigms for Mental Health*, Milan, Italy: Springer International Publishing, 2015.

[27] P. Wu et al., "Objectifying facial expressivity assessment of Parkinson's patients: preliminary study," *Computational and Mathematical Methods in Medicine*, Nov. 2014.

[28] W. Quan, B. J. Matuszewski, and L-K. Shark, "Facial asymmetry analysis based on 3-D dynamic scans," in Proc. *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, Seoul, South Korea: IEEE, 2012.

[29] J. R. Delannoy and T. E. Ward, "A preliminary investigation into the use of machine vision techniques for automating facial paralysis rehabilitation therapy," in Proc. *Signals and Systems Conference (ISSC 2010), IET Irish*, Cork, Ireland: IEEE, 2010, pp. 228-232.

[30] M., Schätz et al., "Face movement analysis with MS Kinect," in Proc. *Computational Intelligence for Multimedia Understanding (IWCIM), 2016 International Workshop on*, Reggio Calabria, Italy: IEEE, 2016.

[31] C. Lanz et al., "Automated Classification of Therapeutic Face Exercises using the Kinect," in Proc. *VISAPP 2013*, Barcelona, Spain, pp. 556-565

[32] C. Dittmar, J. Denzler, and H-M. Gross, "A Feedback Estimation Approach for Therapeutic Facial Training," in Proc. *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, Washington DC, USA: IEEE, 2017.

[33] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in Proc. *IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, 2013. pp. 532-539.

[34] M. Soleymani et al., "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, Jan. 2012, pp. 42-55.

[35] P. Ekman and W. V. Friesen. Manual for the facial action coding system. *Consulting Psychologists Press*, 1978.

[36] Z. S. Nasreddine et al., "The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 4, Apr. 2005, pp. 695-699.

[37] B. R. Brooks et al., "El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis," *Amyotrophic Lateral Sclerosis and other Motor Neuron Disorders*, vol. 1, no. 5, Jan. 2000, pp. 293-299.

[38] J. M. Cedarbaum 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, Oct. 1999, pp. 13-21.

[39] J. R. Duffy, "Motor speech disorders: clues to neurologic diagnosis," In *Parkinson's Disease and Movement Disorders*, C. H. Adler and J. E. Ahlskog: Humana Press, 2000, pp. 35-53.

[40] Y. Yunusova, et al., "A protocol for comprehensive assessment of bulbar dysfunction in amyotrophic lateral sclerosis (ALS)." *Journal of visualized experiments: JoVE*, vol. 48, Feb 2011.

[41] A. Tsanas et al., "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, May 2012, pp. 1264-1271.

[42] C-W. Hsu, C-C. Chang, and C-J. Lin, "A practical guide to support vector classification," 2014 [Online]. Available: http:// www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf [Accessed: September 29, 2017].

[43] J. Wang et al., "Towards Automatic Detection of Amyotrophic Lateral Sclerosis from Speech Acoustic and Articulatory Samples," in Proc. *Interspeech 2016*, San Francisco, CA, USA: ISCA, 2016.

[44] S. Zhu, et al., "Face alignment by coarse-to-fine shape searching." In Proc. *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA: IEEE, 2015, pp. 4998-5006

[45] O. Tuzel, T. K. Marks, and S. Tambe, "Robust face alignment using a mixture of invariant experts, " In Proc. *European Conference on Computer Vision*, Amsterdam, The Nederlands: Springer International Publishing, 2016, pp. 825-841.

[46] MakeHuman – www.makehuman.org. [Accessed: September 29, 2017]