Computer Methods and Programs in Biomedicine 229 (2023) 107309

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Voiceprint and machine learning models for early detection of bulbar dysfunction in ALS☆

Alberto Tena[a,b,*], Francesc Clarià[b], Francesc Solsona[b], Mónica Povedano[c]

[a] CIMNE. Building C1, North Campus, UPC. Gran Capità, Barcelona 08034, Spain
[b] Department of Computer Science and Industrial Engineering, University of Lleida, Lleida, 25001 Spain
[c] Neurology Department, Hospital Universitari de Bellvitge, L'Hospitalet, Barcelona, Spain

ARTICLE INFO

Article history:
Received 6 April 2022
Revised 2 November 2022
Accepted 12 December 2022

Keywords:
ALS
Bulbar dysfunction
Voice
Diagnosis
Machine learning

ABSTRACT

Background and Objective: Bulbar dysfunction is a term used in amyotrophic lateral sclerosis (ALS). It refers to motor neuron disability in the corticobulbar area of the brainstem which leads to a dysfunction of speech and swallowing. One of the earliest symptoms of bulbar dysfunction is voice deterioration characterized by grossly defective articulation, extremely slow laborious speech, marked hypernasality and severe harshness. Recently, research efforts have focused on voice analysis to capture this dysfunction. The main aim of this paper is to provide a new methodology to diagnose this dysfunction automatically at early stages of the disease, earlier than clinicians can do.

Methods: The study focused on the creation of a voiceprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels and the computation of the five principal and independent components of this pattern. Then, a set of statistically significant features was obtained using multivariate analysis of variance and the outcomes of the most common supervised classification models were obtained.

Results: The best model (random forest) obtained an accuracy, sensitivity and specificity of 88.3%, 85.0% and 95.0% respectively when classifying bulbar vs. control participants but the results worsened when classifying bulbar vs. no-bulbar patients (accuracy, sensitivity and specificity of 78.7%, 80.0% and 77.5% respectively for support vector machines). Due to the great uncertainty found in the annotated corpus of the ALS patients without bulbar involvement, we used a safe semi-supervised support vector machine to relabel the ALS participants diagnosed without bulbar involvement as bulbar and no-bulbar. The performance of the results obtained increased, especially when classifying bulbar and no-bulbar patients obtaining an accuracy, sensitivity and specificity of 91.0%, 83.3% and 100.0% respectively for support vector machines. This demonstrates that our model can improve the diagnosis of bulbar dysfunction compared not only with clinicians, but also the methods published to date.

Conclusions: The results obtained demonstrate the efficiency and applicability of the methodology presented in this paper. It may lead to the development of a cheap and easy-to-use tool to identify this dysfunction in early stages of the disease and monitor progress.

1. Introduction

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease characterized by a progressive loss of both upper and lower motor neurons leading to muscular atrophy, paralysis and death. Currently, there is no cure for ALS, although early detection can slow progress [1].

ALS is known as spinal (80%; limb or spinal onset) and bulbar (20%; bulbar onset). The first bulbar symptoms appear early in the disease in bulbar ALS, but can also appear in later stages of spinal

☆ This work was supported by the Intelligent Energy Europe (IEE) programme and the Ministerio de Economía y Competitividad under contract TIN2017-84553-C2-2-R. FS is member of the research group 2017-SGR363, funded by the Generalitat de Catalunya.

* Corresponding author.
E-mail addresses: atd5@alumnes.udl.cat (A. Tena), francesc.claria@udl.cat (F. Clarià), francesc.solsona@udl.cat (F. Solsona), mpovedano@bellvitgehospital.cat (M. Povedano).

https://doi.org/10.1016/j.cmpb.2022.107309

ALS. Early detection of bulbar dysfunction may be the key to effectively increasing patients' survival. However, diagnosing voice disturbances could be challenging due to the limitations of human hearing [2].

Several studies demonstrated that the voice is one of the most important aspects for detecting bulbar dysfunction. Norel et al. [3] developed machine models for recognizing the presence and severity of ALS using a variety of frequency, spectral, and voice quality features. An et al. [4] used Convolutional Neural Networks (CNNs) to classify the intelligible speech produced by patients with ALS compared with healthy individuals. Wang et al. [5] used Support Vector Machines (SVM) and Neuronal Networks (NN) employing acoustic features and adding articulatory motion information (from the tongue and lips). Suhas et al. [6] used SVM and Deep Neuronal Networks (DNNs) based on mel frequency cepstral coefficients (MFCCs) to capture this dysfunction. Recent studies [7–9] demonstrated the feasibility of automatic detection of bulbar dysfunction through phonatory obtained from vowel utterance even before it becomes perceptible to human hearing. Great uncertainty in the annotated corpus of the ALS patients without bulbar involvement was found. In a more recent study [10] time-frequency features were added improving the results. Although all methods performed well in general, this performance dropped significantly when diagnosing bulbar involvement among ALS patients. The aforementioned study argued that the main causes of this uncertainty was a small and wrongly annotated corpus of the ALS patients without bulbar involvement. This suggests that subjective methods employed by clinicians could lead them to misdiagnose this dysfunction. This is coherent with the NEALS bulbar subcommittee, which calls for objective-based approaches.

We conjecture that the diagnosis of ALS patients with bulbar dysfunction would greatly benefit from the creation of a voiceprint able to detect bulbar dysfunction in ALS before the first symptoms can be detected by human hearing. This could be done effectively by means of analyzing a pattern generated from the quasi-periodic waveform produced by the vocal folds when a vowel is elicited. Quasi-periodic waveform analysis has been applied to several clinical applications such as heartbeat detection, cardiopulmonary modeling and intrinsic brain activity detection [11,12]. Furthermore, performance could be increased by correcting the bias as well as enlarging the corpus upsampling it [13], and relabeling bulbar and non-bulbar ALS patients by using semi-supervised classifiers, as pointed out in [14,15].

Our objective (and contribution) is to create a machine-learning model obtained by applying supervised and unsupervised classifiers and upsampling to improve the corpus for diagnosing bulbar dysfunction. This will be done through the creation of a voiceprint consisting of a pattern generated from the quasi-periodic components of a steady portion of the five Spanish vowels, and the five principal and independent components of this pattern. This model should behave properly with small and usually badly annotated corpus, the kind associated with rare diseases (i.e. ALS without bulbar involvement).

2. Methods

The methods presented in this section were implemented and a synthetic dataset based on a random sample of the corpus is freely available online [16].

2.1. Participants

Forty-five ALS participants (26 males and 19 females) aged from 37 to 84 (M = 57.8 years, SD = 11.8 years) and 18 control subjects (9 males and 9 females) aged from 21 to 68 (M = 45.2 years, SD = 12.2 years) took part in this study. All the ALS participants were diagnosed by a neurologist and to participate in the study it was required that they were native Spanish speakers. No restriction was established related to the participants being bilinguals as native speakers of other languages, such as Catalan, which is very common in the Catalonia region where the study occurred.

Bulbar dysfunction was diagnosed by following subjective clinical approaches [17] and the neurologist made the diagnosis of whether an ALS patient had bulbar dysfunction. Among all the ALS participants, 5 reported bulbar onset and 40 spinal onset. However, at the time of the study, 14 of them presented bulbar symptoms.

To summarize, 14 of the 63 participants were ALS patients diagnosed with bulbar dysfunction (3 males and 11 females) aged from 38 to 84 (M = 56.8 years, SD = 12.3), 31 were ALS patients that did not present this dysfunction (23 males and 8 females) aged from 37 to 81 (M = 58.3 years, SD = 11.7) and 18 were control subjects (9 males and 9 females) aged from 21 to 68 (M = 45.2 years, SD = 12.2 years).

2.2. Vowel recording

Sustained samples of the Spanish vowels, a, e, i, o and u, were elicited under medium vocal loudness conditions for 3-4 seconds. The recordings were made in a regular hospital room using an USB EMITA Streaming GXT 252 microphone connected to a laptop at a sampling rate of 44,100 Hz. 32-bit quantization was done using *Audicity*, an open-source application. Each individual phonation was cut out and anonymously labeled. The boundaries of the speech segments were determined with an oscillogram and a spectrogram using the Praat manual [18], and were audibly checked. The starting point of the boundaries was established at the onset of the periodic energy in the waveform observed in the oscillogram and checked by the appearance of the formants in the spectrogram. The endpoint was established at the end of the periodic oscillation when a marked decrease in amplitude in the periodic energy was observed. It was also identified by the disappearance of the waveform in the oscillogram and the formants in the spectrogram.

2.3. Generating the pattern of the quasi-periodic components of the five Spanish vowels

A sample of 250 ms of each vowel was considered for analysis by taking the middle point at the center of the phonation. This fragment of the signal was normalized by centering each sample to have a mean of 0 and scaled to have a standard deviation of 1 ($x(n)$). A pattern generator was developed to obtain a pattern sequence of the quasi-periodic components of the fundamental frequency of $x(n)$ inspired by [19]. This process consisted of 3 steps.

2.3.1. Detrending method

The baseline wandering of $x(n)$, which is a low-frequency artefact present in signal recordings, was removed by implementing a detrending method. To obtain the trend, a six-order low-pass Butterworth filter with a cutoff frequency of 0.0035 Hz was applied twice (forward and backward) to $x(n)$. The combined filter had zero phase distortion, a filter transfer function equal to the squared magnitude of the implemented Butterworth filter transfer function, and a filter order that was double the order of the Butterworth filter. Then, the detrending signal $x_d(n)$ was obtained by removing the trend from $x(n)$. Finally, each sample of $x_d(n)$ was centered to have a mean of 0. Figure 1a shows $x(n)$ and the trend of $x(n)$ and Fig. 1b shows $x(n)$ and $x_d(n)$.

2.3.2. Marking the quasi-periodic components of $x(n)$

The spectral density $|X_d(f)|^2$ of $x_d(n)$ (Fig. 2) was obtained by means of the discrete Fourier transform (DFT) implementing

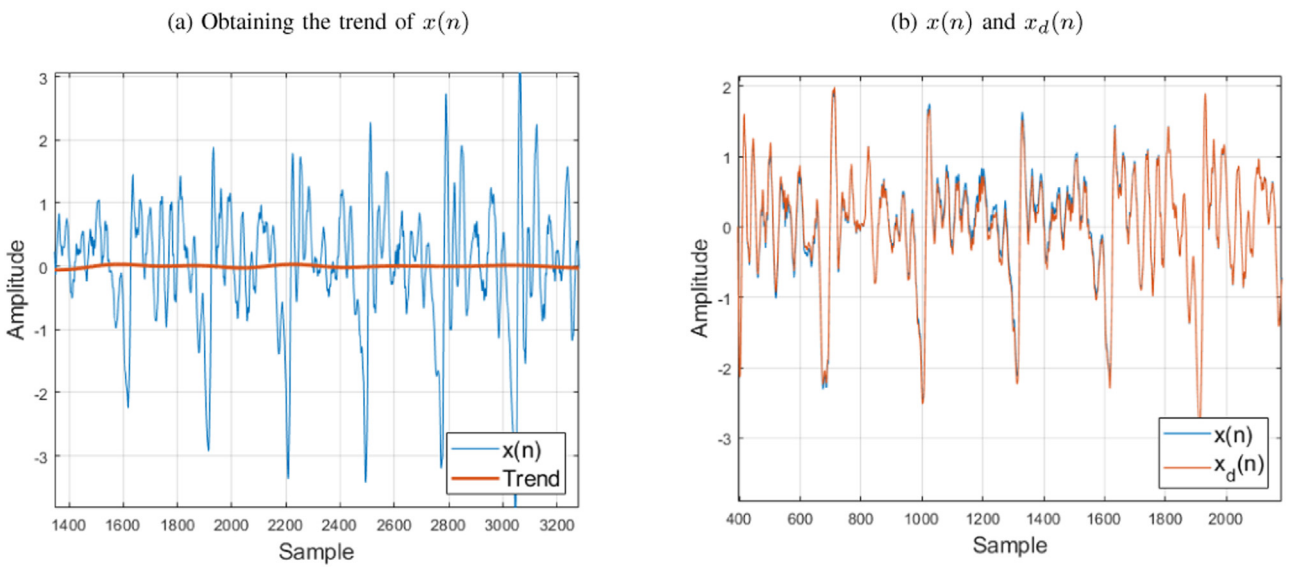


Fig. 1. Detrending method: Removing the trend from $x(n)$.

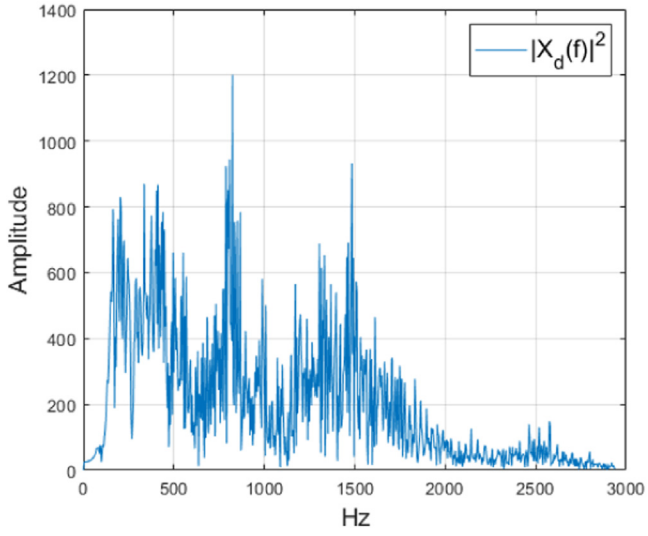


Fig. 2. The spectral density of $x_d(n)$.

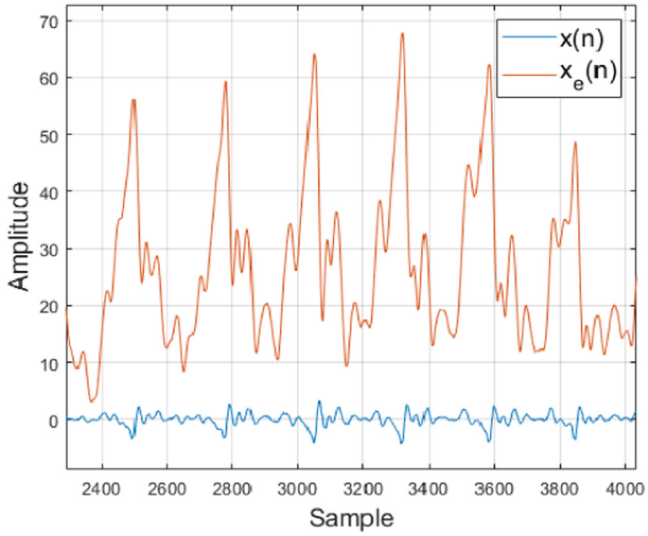


Fig. 3. The signal envelope of $x(n)$.

the fast Fourier transform (FFT) algorithm. To identify the quasi-periods, the samples of $|X_d(f)|^2$ whose frequency was ≥ 300 Hz were considered to identify the peaks of the spectral density. To avoid noise, only the three highest peaks were selected. Finally, the quasi-period of $x_d(n)$ was defined as the lower spectral component of these three peaks (f_r). The number of samples of each quasi-period (n_{rep}) was calculated as the nearest integer of (f_s/f_r), f_s being the recording sampling rate.

The signal envelop $x_e(n)$ was obtained by computing the cumulative sum of $x_d(n)$ and then calculating the envelope of the analytical signal. Figure 3 shows $x_e(n)$ and $x(n)$.

To detect the starting and ending point of each quasi-period, a quasi-sinusoidal signal, $s(n)$, synchronized with the period of $x(n)$ was computed. It was obtained by applying a second-order Butterworth pass-band filter forward and backward to $x_e(n)$ with a cut-off frequency $f_c = f_s/n_{rep}$ Hz. From $s(n)$, a quadratic-bipolar signal ($q(n)$) was generated assigning a constant -A in those samples where $s(n) < 0$, and A in those where $s(n) > 0$. Thus, by differentiating $q(n)$, the zero crossings of the synchronized signal $s(n)$ were obtained, which represent the beginning and end of each quasi period of $x(n)$. Figure 4 illustrates this process. Figure 4a represents $x(n)$, $s(n)$ and $q(n)$ and Fig. 4b depicts the starting and ending points detected of each quasi-period of $x(n)$.

Finally, the pattern function $p(T)$ was obtained as the average of the quasi-periods of $x(n)$, T being the average of the number of samples of the quasi-period of $x(n)$.

2.3.3. Pattern refinement

$p(T)$ was compared with $x_d(n)$ to improve the boundaries of each quasi-period. First, as an adapted filter, the pattern $p(T)$ was inverted and the resulting signal was convolved with $x_d(n)$ to detect the positions of $p(T)$ in $x_d(n)$. The positive values of the resulting signal were taken and the negative values were set at 0.

Each quasi-period detected previously was centered in the position where the maximums values of the convolution were found. The refined pattern, $p_{ref}(T)$ (Fig. 5a) was computed as the average of the quasi-periods of $x_d(n)$ with their new boundaries established.

Finally, $p_{ref}(T)$ was normalized to 550 samples and then decimated to 110 samples to obtain patterns, $p_N(T)$ (Fig. 5b), with the same length.

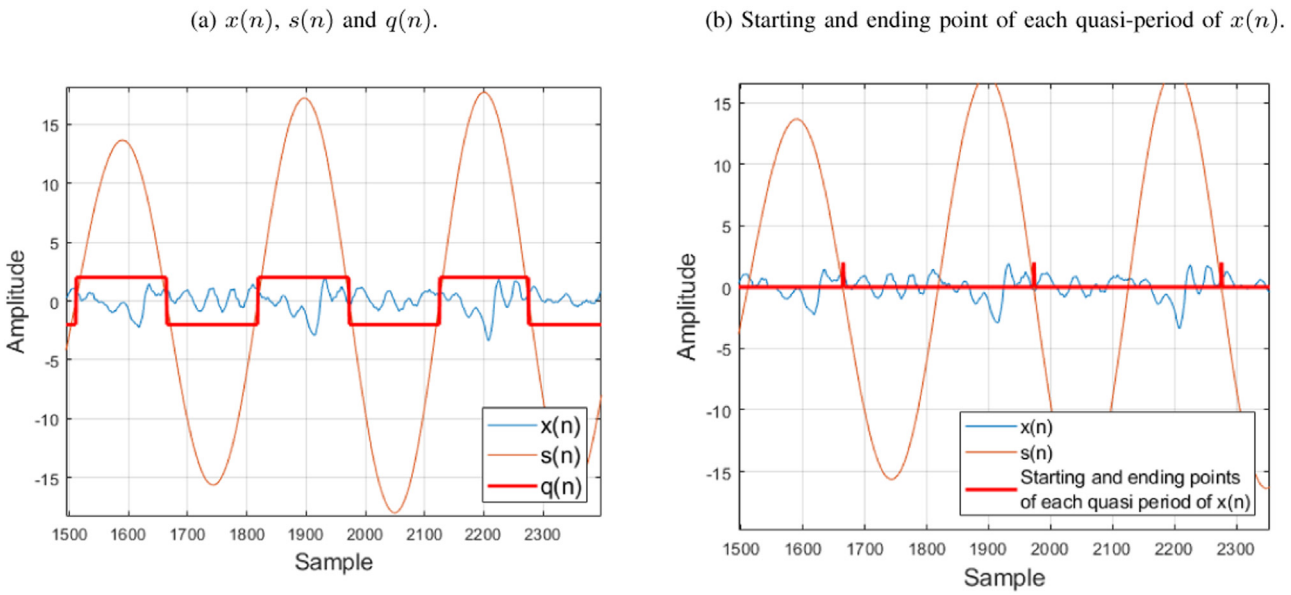


Fig. 4. Detecting the starting and ending point of each quasi-period of $x(n)$.

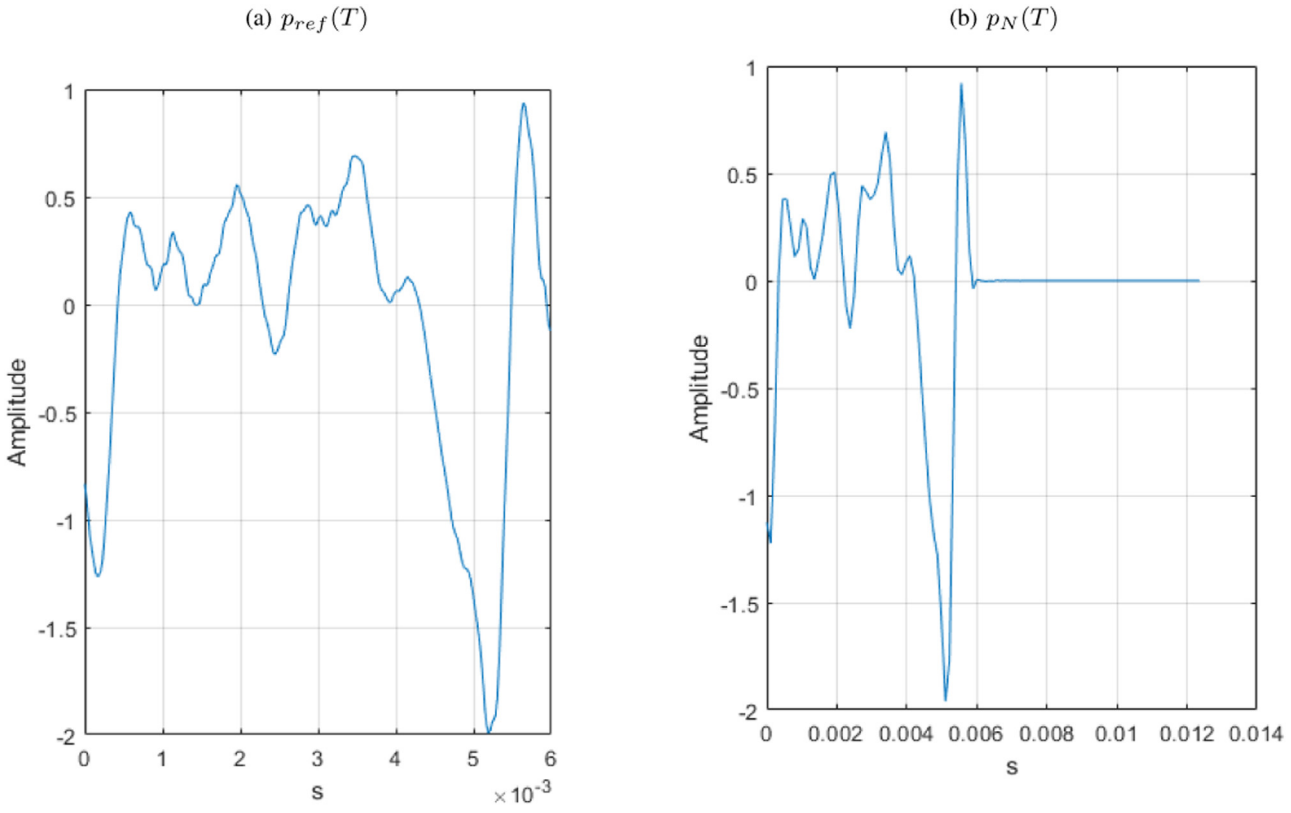


Fig. 5. Pattern refinement and normalization.

2.4. Principal and independent component analysis

Principal Component Analysis (PCA) and Independent Component Analysis (ICA) have great potential in the treatment of medical signals [20]. PCA is a classical technique in statistical data analysis, feature extraction and data reduction, aiming at explaining observed signals as a linear combination of orthogonal principal components. ICA is a technique for array processing and data analysis, aiming at recovering unobserved signals from observed mixtures, exploiting only the assumption of mutual independence between the signals.

In the PCA of $p_N(T)$, five Principal Components (PCs) were computed. The decomposition was obtained as $X = USV^\top$ where X is $p_N(T)$ standardized, U is a unitary matrix and S is the diagonal matrix of singular values s_i. PCs were given by US and V contained the directions in this space that capture the maximal variance of the matrix X.

In the ICA of $p_N(T)$, five independent components (ICs) were extracted by means of a reconstruction independent component analysis (RICA) algorithm [21]. $p_N(T)$ was standardized to have zero mean and identity co-variance. The model $x = \mu + As$ is made up of the five rows of matrix x representing the patterns of the

five vowels with 110 samples for each pattern. μ is a constant represented by a column vector of five rows and s is a matrix in which each row (5) is an independent component with 110 samples. A(5x5) is the mixing matrix. Once the model has been obtained, the five independent components computed were used for the analysis.

2.5. Features obtained for analysis

A total of 70 features were obtained as follow:

- Three entropy measurements per each vowel were obtained by means of the Shannon entropy:
  - From the probability density function of a signal, formed by the pattern $p_N(T)$ repeated the number of periods of $x(n)$ and quantified with $N = 2^q$ levels and $q = 8$. This measurement was coded as entPat1...entPat5.
  - The density function of the five PCs was normalized and quantified with $N = 2^q$ levels and $q = 6$. Then, the Shannon entropy of the probability density function of the five PCs was obtained and coded as entPC1...entPC5.
  - Similarly, to compute the Shannon entropy of the five ICs, the density function of the five ICs was normalized and quantified with $N = 2^q$ levels and $q = 6$ and the Shannon entropy of each IC was obtained. The results were coded as entIC1...entIC5.
- The variance of a signal formed by the pattern $p_N(T)$ repeated the number of periods of $x(n)$ was computed and coded for each vowel as var1...var5.
- The Kurtosis is defined as a measure of outlier-prone. It is calculated from the distribution of a signal formed by the pattern $p_N(T)$ repeating the number of periods of $x(n)$, and coded as kurt1...kurt5. The bias-corrected equation defined in [21] was applied to obtain the Kurtosis.
- The rhythm variability, $RR(n)$, of $x(n)$ was computed by firstly calculating the differences (in seconds) between the quasi-periodic elements of $x(n)$ and dividing the result by the sampling frequency (44.100 Hz). Finally, $RR(n)$ was obtained by reducing the sampling to $fr = 350$. Thus, $RR(n)$ is a signal resampled to $fr = 350$ ($Tr = 0.0029$ s) with a bandwidth of 175 Hz. The mean and the standard deviation of RR ($mean\_RR$ and std_RR) were then computed and coded for each vowel as medRR1...medRR5 and dsvRR1...dsvRR5 respectively.
- The spectrum of $P\_N(f)$ was obtained from the positive and normalized part of the FFT of the autocorrelation of $p_N(T)$. The mean frequency of $P\_N(f)$ ($fmEsPat$) was computed according to Eq. (1) in the frequency band 0 Hz to 2205 Hz and coded for each vowel as fmEsPat1...fmEsPat5.

$$fmEsPat = \int fP\_N(f)\, df \tag{1}$$

- Similarly, the mean frequency of the probability density function of the five PCs was computed and coded as fmEsPC1...fmEsPC5.
- Finally, the average spectral energy was calculated as the integral of $P\_N(f)$. The average spectral energy was computed and normalized to 1 for 5 frequency bands of the total spectrum (0-4,410 Hz): 1, 0-250 Hz; 2, 250-750 Hz; 3, 750-1500 Hz; 4, 1500-2500 Hz; 5, 2500-4,410 Hz. These measurements for the five patterns and the five bands of each pattern were coded for each vowel as enBnEs_a1...enBnEs_a5 and enBnEs_u1...enBnEs_u1, respectively.

2.6. Classification models

Five supervised classification models were implemented in R to measure the classification performance. These models are Random Forest (RF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Neural Networks (NN) and Support Vector Machine (SVM). The classification models were fitted with the features selected. These were standardized by subtracting the mean and centered at 0. 10-fold cross-validation was implemented in R using the caret package [22] to draw suitable conclusions. The upsampling technique with replacement was applied to the training data by making the group distributions equal to deal with the unbalanced dataset that could bias the classification models [13]. Supervised models with classification thresholds of 50% were built. The classification threshold is a value that converts the result of a quantitative test into a simple binary decision by treating the values above or equal to the threshold as positive, and those below as negative.

In addition, the semi-supervised classification model S4VM was implemented using the RSSL package [23]. S4VM returns predicted labels for unlabeled instances. It randomly generates multiple low-density separators and merges their predictions by solving a linear programming problem meant to penalize the cost of decreasing the performance of the classifier, compared to the supervised SVM [24]. As for SVM, a linear kernel was used, and the regularization parameter C for labeled and unlabeled data was set at 0.05.

2.7. Feature selection

To select a subset of relevant features for use in the construction of the classification model, the Multivariant Analysis of Variance (MANOVA), which uses the covariance between the features in testing the statistical significance of the mean differences, was implemented in IBM SPSS Statistics. By using this procedure, it was possible to contrast the null hypothesis in the features obtained.

To perform this statistical analysis, it was assumed that the features had a multivariable normal distribution and no assumptions were made regarding the homogeneity of the variance or the correlation between the features. A significance value of $p < 0.05$ was considered sufficient to assume the existence of feature differences between the four groups analyzed.

2.8. Experiments

The participants in this study belonged to three different groups: the control group with 18 participants, labeled as C, the group with 14 ALS participants diagnosed with bulbar dysfunction, labeled B, and the group with 31 ALS participants not diagnosed with bulbar dysfunction, labeled NB. In addition, the A label was added to every ALS participant, with or without bulbar dysfunction.

Three experiments were performed with these groups:

1. Performance evaluation of the supervised models for 4 cases (C vs. B, C vs. NB, B vs. NB and C vs. A) by using the original corpus.
2. Re-labeling of the NB participants as B' and C' by applying the semi-supervised S4VM algorithm. Thus, the NB group was removed.
3. Re-evaluation of the model performance with four new groups of participants: C vs. B+ (B + B'), C vs. NB- (NB - B'), B+ vs. NB- and C+ (C + C') vs. B+.

The first experiment obtained the outcomes of the models using the original corpus. Next, due to the great uncertainty found in the ALS participants diagnosed without bulbar involvement [2,9], it was intended to re-label the participants of the NB group as B and C using the semi-supervised S4VM model. We tried to obtain a new corpus that contains elements classified as bulbar (B') or control (C') by S4VM among those who were previously diagnosed as non-bulbar by a clinician (NB). In the third experiment, the models outcomes were again obtained by changing the composition of the

B and NB groups by adding B' to B (B+) and removing C' from NB (NB-).

2.9. Performance metrics

There are several metrics for evaluating classification algorithms [25]. The Accuracy, Sensitivity and Specificity metrics, the most popular ones, were used to evaluate the performance of the classification models.

3. Results

Firstly, the voiceprint representations and the features selected in relation to the four cases (C vs B, C vs NB, B vs NB and C vs A) are presented. Then, the performance of the classification models is evaluated.

3.1. Voiceprint representations to detect bulbar dysfunction in ALS

The voiceprint for detecting bulbar dysfunction in ALS consisted of the computation of $p_N(T)$, the 5 PCs of $p_N(T)$ and the 5 ICs of $p_N(T)$ for the 5 Spanish vowels.

Figure 6 shows the voiceprint computed of a ALS patient. Figure 6(a) shows the $p_N(T)$ of the five Spanish Vowels (a, e, i, o, u). Figure 6(b) shows the 5 PCs of $p_N(T)$ of the five Spanish Vowels. Figure 6(c) shows the 5 ICs of $p_N(T)$ of the five Spanish Vowels. Figure 6(d) shows the spectrum of $p_N(T)$ of the five Spanish Vowels. Figure 6(e) shows the spectrum of the 5 PCs of $p_N(T)$ of the five Spanish Vowels. Figure 6(f) shows the probability density function of the 5 ICs of $p_N(T)$ of the five Spanish Vowels.

3.2. Features selected

A total of 70 features were obtained. The MANOVA analysis was applied to select the statistically significant features (p-value<0.05) for the four comparisons analyzed: C vs. B, C vs. NB, B vs. NB and C vs. A. The features not showing statistical significance (p-value≥ 0.05) were discarded. The box plots of the statistically significant features are depicted in Fig. 7.

In the case C vs B, a set of 19 statistically significant features (p-value<0.05) were obtained. These were medRR2, medRR3, medRR5, fmEsPat1, enBnEs_a3, enBnEs_e4, enBnEs_e5, enBnEs_i5, enBnEs_o5, entPat2, entPat3, entPat4, entPC1, entPC2, entPC4, entPC5, entIC2, entIC3 and entIC4.

In the case C vs NB, a set of 2 statistically significant features were obtained. These were enBnEs_o4 and enBnEs_o5.

In the case B vs NB, a set of 20 statistically significant features were obtained. These were medRR1, medRR2, medRR3, medRR4, enBnEs_e4, enBnEs_e5, enBnEs_i3, entPat1, entPat2, entPat4, entPC1, entPC2, entPC3, entPC4, entPC5, entIC1, entIC2, entIC3, entIC4 and entIC5.

In the case C vs A, a set of 3 statistically significant features were obtained. These were fmEsPat1, enBnEs_o4 and enBnEs_o5.

3.3. Classification model performance and experiments

In the first experiment, the classification models were fitted with the features selected per each case. Table 1 shows the classification performance (*Accuracy, Sensitivity* and *Specificity* metrics) of the classification models tested for the four cases defined with the original labels (B, NB and C).

In the case C vs. B, the results indicate that RF and SVM have a good classification performance. RF obtained the best *Accuracy*, 88.3%, with a *Sensitivity* of 85.0% and a *Specificity* of 95.0%. SVM obtained an *Accuracy* of 86.5% with a *Sensitivity* of 88.3% and a *Specificity* of 85.0%. NN, LR and LDA showed a poorer performance, obtaining respective *Accuracies* of 66.7%, 56.7% and 54.2% respectively.

Table 1
Model performance for the first experiment.

		C vs B	C vs NB	B vs NB	C vs A
RF	Accuracy	**88.3**	48.7	66.7	68.6
	Sensitivity	85.0	46.7	40.0	**84.0**
	Specificity	**95.0**	50.0	75.8	30.0
LR	Accuracy	56.7	62.0	60.2	65.0
	Sensitivity	45.0	68.3	60.0	66.0
	Specificity	65.0	50.0	59.2	**65.0**
LDA	Accuracy	54.2	62.0	54.0	65.2
	Sensitivity	45.0	68.3	60.0	68.5
	Specificity	60.0	50.0	51.7	60.0
NN	Accuracy	66.7	59.2	66.3	60.5
	Sensitivity	70.0	63.3	60.0	64.0
	Specificity	65.0	50.0	68.3	55.0
SVM	Accuracy	86.5	**68.0**	**78.7**	**73.1**
	Sensitivity	**88.3**	**71.7**	**80.0**	79.0
	Specificity	85.0	**60.0**	**77.5**	60.0

Table 2
Model performance for the third experiment.

		C vs B+	C vs NB-	B+ vs NB-	C+ vs B+
RF	Accuracy	**93.5**	69.3	89.7	**92.4**
	Sensitivity	**96.6**	66.7	**83.3**	83.3
	Specificity	**90.0**	**70.0**	96.7	**97.5**
LR	Accuracy	56.0	69.2	71.7	71.4
	Sensitivity	53.3	73.3	76.7	66.7
	Specificity	60.0	60.0	65.0	75.0
LDA	Accuracy	62.9	69.2	82.3	80.9
	Sensitivity	60.0	73.3	76.7	78.3
	Specificity	65.0	60.0	88.3	82.5
NN	Accuracy	75.9	68.7	86.9	82.6
	Sensitivity	80.0	**80.0**	**83.3**	83.3
	Specificity	70.0	50.0	91.6	82.5
SVM	Accuracy	89.2	**69.6**	**91.0**	92.1
	Sensitivity	90.1	73.3	**83.3**	**86.7**
	Specificity	**90.0**	60.0	**100**	95.0

In the cases C vs. NB, B vs. NB and C vs. A, poorer results were obtained. In all these cases SVM obtained the best *Accuracy*, these being 68.0%, 78.7% and 73.1% respectively.

In the second experiment, S4VM was applied from the data labeled C and B to estimate the class of NBs which were split into C' and B'. From the total of 31 NBs, 9 were split as B' and 22 as C'.

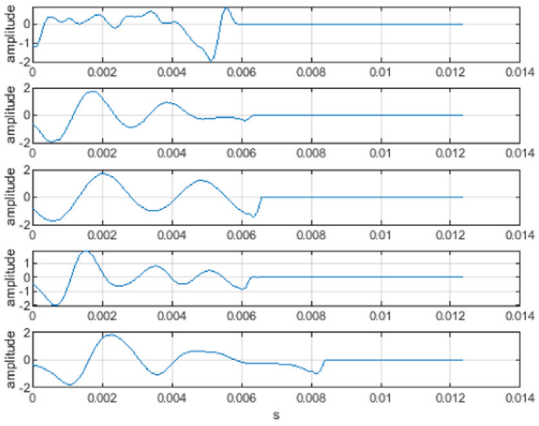
In the third experiment, the classification models were fitted with the features selected and tested for the four new cases (C vs. B+, C vs. NB-, B+ vs. NB- and C+ vs. B+). Table 2 shows the classification performance (*Accuracy, Sensitivity* and *Specificity* metrics) for this experiment.

In the case of C vs. B+, the results indicate that RF and SVM have good classification performance. RF obtained the best *Accuracy*, 93.5%, with a *Sensitivity* of 96.6% and a *Specificity* of 90.0%. SVM obtained an *Accuracy* of 89.2% with a *Sensitivity* of 90.1% and a *Specificity* of 90.0%. NN obtained an *Accuracy* of 75.9% with a *Sensitivity* of 80.0% and a *Specificity* of 70.0%. LR and LDA showed poorer performance, obtaining *Accuracies* of 56.0% and 62.9% respectively.

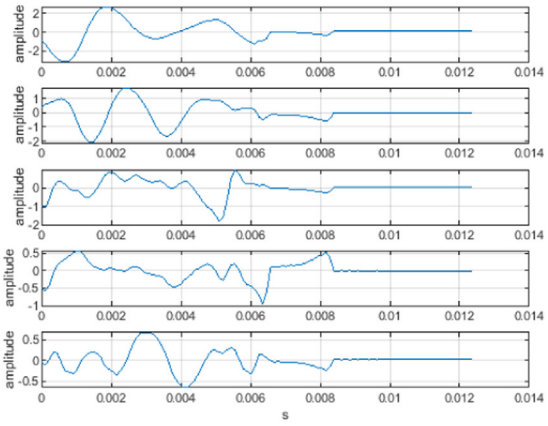
In the B+ vs. NB- and C+ vs. B+ cases, good model classification performance was also observed.

In B+ vs. NB-, SVM obtained the best *Accuracy*, 91.0%, with a *Sensitivity* of 83.3% and a *Specificity* of 100.0%. RF obtained an *Accuracy* of 89.7% with a *Sensitivity* of 83.3% and a *Specificity* of 96.7%. NN obtained an *Accuracy* of 86.9% with a *Sensitivity* of 83.3% and a *Specificity* of 91.6%. LDA obtained an *Accuracy* of 82.3% with a *Sensitivity* of 76.7% and a *Specificity* of 88.3%. Finally, LR performed the

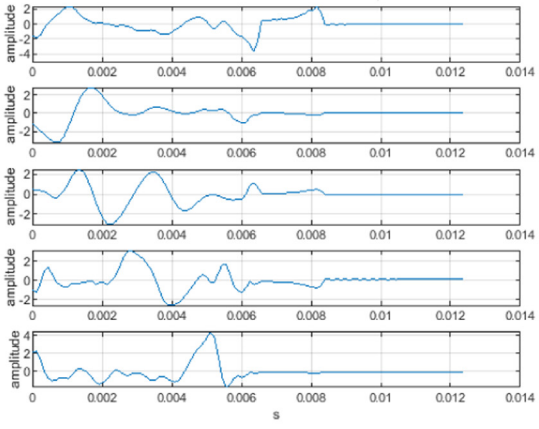
(a) $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom)



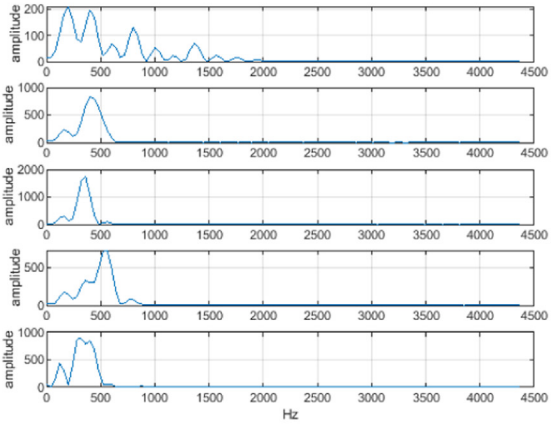
(b) Principal Components of $p_N(T)$ of the five Spanish vowels ordered from the highest (PC1 on the top) to the lowest contribution (PC5 on the bottom)



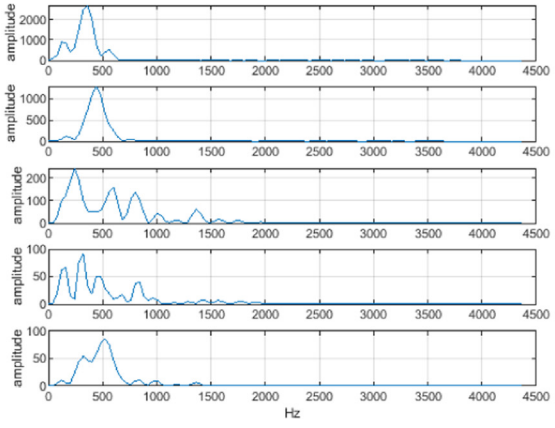
(c) Independent Components of $p_N(T)$ of the five Spanish vowels ordered from the highest (IC1 on the top) to the lowest contribution (IC5 on the bottom)



(d) Spectrum of $p_N(T)$ of the five Spanish vowels: a (top), e, i, o, u (bottom)



(e) Spectrum of the Principal Components, PC1 to PC5, of $p_N(T)$ of the five Spanish vowels ordered from the highest to lowest contribution



(f) Probability Density Function of the Independent Components of $p_N(T)$ of the five Spanish vowels

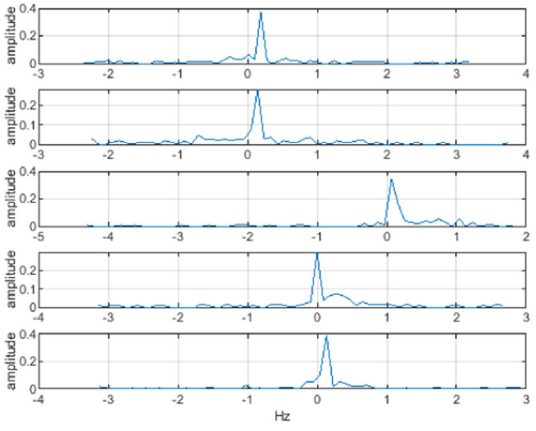


Fig. 6. Voiceprint for a patient.

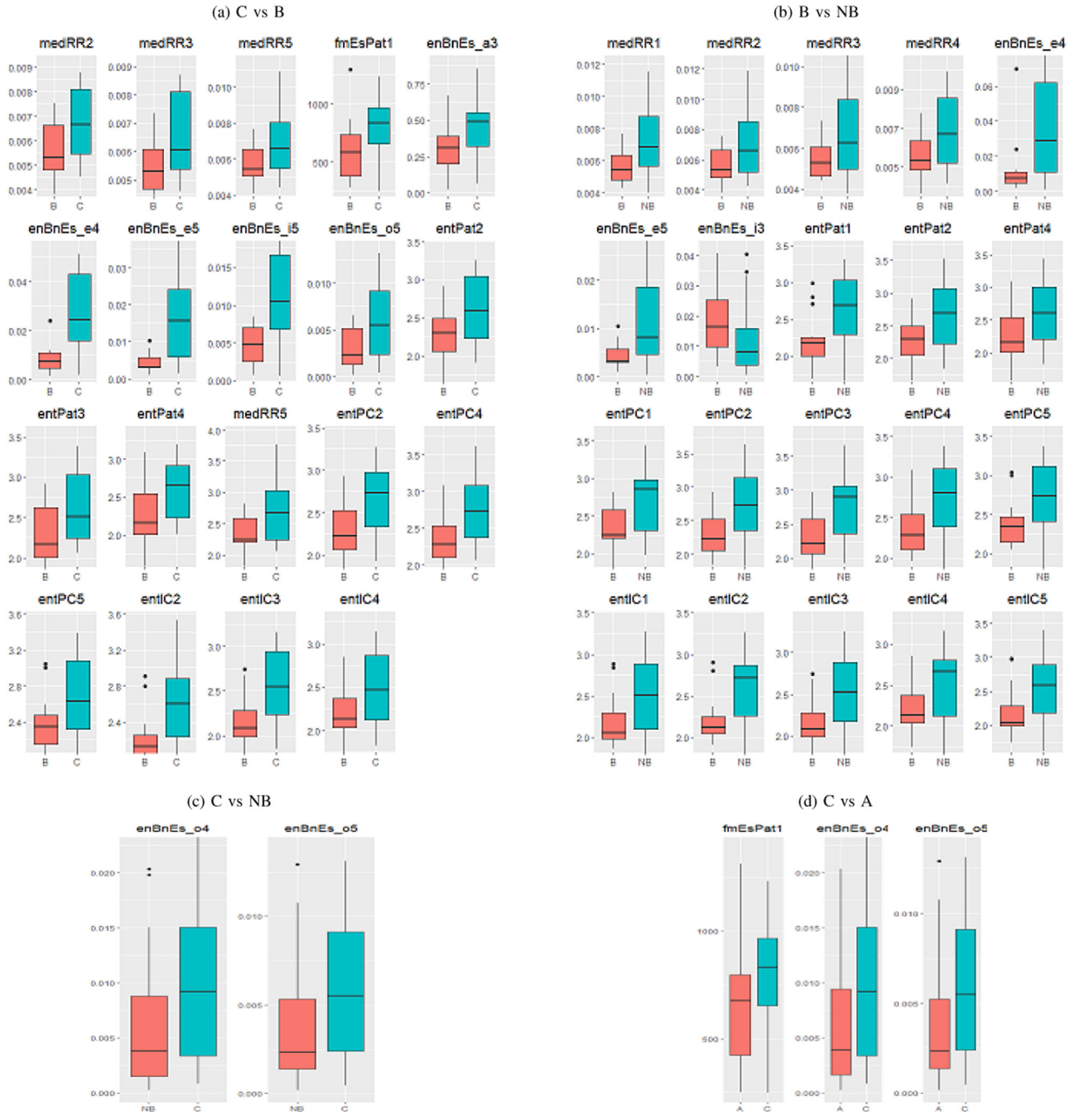


Fig. 7. Box plots of the statistically significant features per each case.

worst performance with an *Accuracy*, 71.7%, a *Sensitivity* of 76.7% and a *Specificity* of 65.0%.

In C+ vs. B+, RF obtained the best *Accuracy*, 92.4%, with a *Sensitivity* of 83.3% and a *Specificity* of 97.5%. SVM obtained an *Accuracy* of 92.1% with a *Sensitivity* of 86.7% and a *Specificity* of 95.0%. NN obtained an *Accuracy* of 82.6% with a *Sensitivity* of 83.3% and a *Specificity* of 82.5%. LDA obtained an *Accuracy* of 80.9% with a *Sensitivity* of 78.3% and a *Specificity* of 82.5%. Finally, LR had the worst performance with an *Accuracy* of 71.4%, a *Sensitivity* of 66.7% and a *Specificity* of 75.0%.

Finally, in C vs NB- poorer results were obtained. All models showed similar performance. SVM, RF, LR, LDA and NN obtained *Accuracies* of 69.6%, 69.3%, 69.2%, 69.2% and 68.7%.

4. Discussion

4.1. Principal findings

We have carried out a preliminary assessment of the potential for obtaining a voiceprint for an early detection of bulbar dysfunction in ALS patients. This was motivated by the need for a standardized diagnostic procedure for assessing bulbar dysfunction and new methodologies based on objective measurements [2].

The study demonstrated the feasibility of the methodology proposed. Its major benefit is to provide a methodology based on objective measures to identify bulbar dysfunction in early stages of the ALS disease. We suggest two new labels, C' and B', to improve

the diagnosis of those patients in whom bulbar dysfunction has not yet been detected by the current subjective procedures.

From the voiceprint, a total of 70 features were obtained. These were: entPat1...entPat5, entPC1...entPC5, entIC...entIC5, var1...var5, kurt1...kurt5, medRR1...medRR5, dsvRR1...dsvRR5, fmEsPat1...fmEsPat5, fmEsPC1...fmEsPC5, enBnEs_a1...enBnEs_a5, enBnEs_e1...enBnEs_e5, enBnEs_i1...enBnEs_i5, enBnEs_o1... enBnEs_o5 and enBnEs_u1...enBnEs_u5.

The first experiment showed the performance of the machine learning models used for the four cases. From the good results achieved by C vs. B, it can be inferred that the methodology proposed is good at detecting bulbar dysfunction, RF and SVM being the best models for performing this task. The poor performance obtained in C vs. NB revealed a similar voice performance of Cs and NBs, as expected. Instead, the performance obtained in B vs. NB may indicate that some NBs voices could be affected in some NBs but this may yet not be perceptible to human hearing.

The second experiment revealed that 9 of the total of 31 NBs may have bulbar dysfunction. This result is consistent with the previous statement that indicated that the voices of some NBs could be affected. We suggest labelling these patients as B' if their voices show a similar performance to Bs and C' if they are similar to C.

The third experiment performed better than the first one when B' and C' labels were considered. In C vs B+, RF outperformed the results obtained in the first experiment. Similarly, in B+ vs. NB-, the classification performance was greatly improved.

In general, the third experiment achieved better performance than the first one. This indicates that our method can diagnose bulbar dysfunction better than clinicians with the current subjective approaches.

4.2. Comparison with prior work

This study is consistent with Tena et al. [9], which found a great uncertainty in ALS patients in whom bulbar dysfunction was not detected yet suggesting that some of them were misdiagnosed. It is also consistent with Plowman et al. [2], which indicated the difficulties in diagnosing bulbar dysfunction by subjective approaches. In many cases, the perturbance in those subjects' voices could not be appreciated by the human ear until advanced stages of the disease. We went a step further by providing two new labels, B' and C', to achieve an earlier and more accurate diagnosis.

This study is also in line with Suhas et al., Vashkevich et al., Norel et al., An et al. and Wang et al. [3–6,8], who demonstrated that voice is one of the most important aspects for detecting bulbar dysfunction. Suhas et al. obtained accuracies of 92.2% using SVM and DNNs based on MFCCs. Vashkevich et al. achieved 90.7% accuracy (sensitivity 86.7%, specificity 92.2%) with LDA using perturbation measurements. Norel et al. identified acoustic features in naturalistic contexts, achieving 83% accuracy (sensitivity 86%, specificity 78%) using SVM. An et al. implemented CNNs to classify the intelligible speech produced by patients with ALS and healthy individuals. The experimental results indicated a sensitivity of 76.9% and a specificity of 92.3%. Wang et al. [5] implemented SVM and NN using acoustic features and adding articulatory motion information (from tongue and lips). Accuracies of 91.7% were obtained using only acoustic features, increasing to 96.5% with the addition of both lip and tongue data. Adding motion measures increased the classifier accuracy significantly at the expense of including more invasive measurements to obtain the data.

These studies only focused on B vs. C cases. We investigated the means of optimizing accuracy in detecting ALS bulbar dysfunction by only analyzing the voices of patients.

To date, only Tena et al. [9,10] have conducted studies considering additional cases. In [9], they used phonatory subsystem features, such as jitter, shimmer, harmonic-to-noise ratio and pitch and PCA. In [10] a set of time-frequency features were added to the analysis and better results were achieved. The performance of several machine learning models were evaluated considering four scenarios (C vs. B, C vs. NB, B vs. NB and C vs. A). In C vs. B, they obtained an *Accuracy* of 98.1% with a *Sensitivity* of 96.6% and *Specificity* of 100.0% using SVM. The results worsened in B vs. NB (*Accuracy* of 84.8% with *Sensitivity* of 92.3% and *Specificity* of 75.0% using RF). In C vs. NB and C vs. A, good results were also obtained, RF being the model which performed best in both cases (*Accuracies* of 94.1% and 95.8% respectively). In this study, in C vs. B, an *Accuracy* of 88.3% was obtained for RF with a *Sensitivity* of 85.0% and a *Specificity* of 95.0%. This performance improved when considering B' patients, C vs. B+, obtaining an *Accuracy* of 93.5% outperforming the results of [3–5]. In B vs. NB, we obtained an *Accuracy* of 78.7% (SVM). This performance was greatly improved when considering B' patients, B+ vs. NB-, obtaining an *Accuracy* of 91.0% with a *Sensitivity* of 83.3% and a *Specificity* of 100.0% outperforming the results obtained by [9,10]. This suggests that having well-annotated patients is essential for properly assessing bulbar dysfunction in B vs. NB. We demonstrated that semi-supervised classification models such as S4VM are useful tools for performing this task.

4.3. Limitations

The size and bias of this study is heavily influenced by the fact that ALS is a rare and a very heterogeneous disease where not all the patients present the same symptomatology. Although upsampling and semi-supervised classifier techniques were used to correct the bias, it would be necessary to increase the number of participants to draw definitive conclusions.

However, we proved that the method presented can be successfully applied to such a corpus. The question is what the outcomes should be when applying it to a large enough corpus. The outcomes indicate that accuracy could increase much more. A specific study should be performed to determine the extent of this increase.

Furthermore, detecting bulbar dysfunction inevitably reflects the clinician's own learned approach, which must depend on a similar comparative data set analysis, gained by so-called clinical experience. As such there are hidden variables; for example, dental health or absent teeth, glossal disorders, respiratory dysfunction altering breath holding and rhythm of breathing, nasal obstruction, nasal-sinus mucus discharge, etc. When a machine learning paradigm is applied to detecting bulbar dysfunction, but it also applies to the majority of the clinical problems, these hidden variables could lead to detecting bulbar dysfunction when, actually, the speech disorder is produced by alternative disorders. This has an effect in specificity because subjects affected by these alternative disorders could have been predicted as suffering bulbar dysfunction when actually they would not be suffering from the disease. Although this is a limitation, the ALS subjects who participated in this study were diagnosed and selected by a neurologist, bounding the possible effects of this phenomenon.

Additionally, concerning the acoustic analysis employed for the detection of the voice impairment, neurologists are accustomed to using certain difficult phrases when testing speech quality, and these include requirements for rhythmic cadence, constancy of utterance for long sounds, repetitive consonantly driving utterances, as well as simple vowels. In this study, the five Spanish vowels were analysed rather than any consonants or other sounds existing in the Spanish language. Therefore, the analysis is not standardized to the extent that it is not in line with a classical clinical analysis. Additional analysis are envisaged to analyse other Spanish existing sounds to obtain the more accurate voiceprint as possible.

4.4. Conclusions

Promising outcomes in detecting bulbar dysfunction were obtained when comparing ALS patients with and without this dysfunction in early stages of the disease, or prior to being diagnosed by clinicians.

This could lead to the development of a cheap and simple tool that may help to develop standardized diagnostic procedures for assessing bulbar dysfunction based on objective measures and monitor progress. This directly addresses a recent statement released by the NEALS bulbar subcommittee regarding the need for objective-based approaches [2].

Due to the great uncertainty of the corpus, we highlight the importance of improving the annotation of ALS patients as regards bulbar dysfunction to develop powerful machine learning models able to distinguish this dysfunction. We provide two new labels, C' and B', and demonstrate that Semi-Supervised Machine learning models could help in the early detection of this dysfunction. Yet, further analyses are needed to develop this concept fully. These include performing longitudinal studies in which patients' diagnosis are retrieved at several follow-ups.

The usefulness of this methodology is that it could be applied to the automated identification and early diagnosis of many other neurological or respiratory illnesses where obtaining a large enough and well-annotated corpus is difficult.

Declaration of Competing Interest

Authors declare no conflicts of interest.

Acknowledgments

This work was approved by the Research Ethics Committee for Biomedical Research Projects (CEIm) at the Bellvitge University Hospital in Barcelona and was supported by the Ministerio de Economía y Competitividad (TIN2017-84553-C2-2-R) and the Ministerio de Ciencia e Innovacion (PID2020-113614RB-C22). AT is a member of CIMNE, a Severo Ochoa Centre of Excellence (2019-2023) under grant CEX2018-000797-S, funded by MCIN/AEI/10.13039/501100011033. The Neurology Department of the Bellvitge University Hospital in Barcelona permitted the recording of the voices of the participants in its facilities. The clinical records were provided by Carlos Augusto Salazar Talavera. Dr. Marta Fulla and Maria Carmen Majos Bellmunt contributed advice about the process of eliciting the sounds.

References

[1] C. Carmona-Duarte, et al., Study of several parameters for the detection of amyotrophic lateral sclerosis from articulatory movement, Loquens 4 (2017), doi:10.3989/loquens.2017.038.

[2] E.K. Plowman, L.C. Tabor, J. Wymer, G. Pattee, The evaluation of bulbar dysfunction in amyotrophic lateral sclerosis: survey of clinical practice patterns in the United States, Amyotroph. Lateral Scler. Frontotemporal Degener. 18 (5–6) (2017) 351–357, doi:10.1080/21678421.2017.1313868.

[3] R. Norel, M. Pietrowicz, C. Agurto, S. Rishoni, G. Cecchi, Detection of amyotrophic lateral sclerosis (ALS) via acoustic analysis, bioRxiv (2018), doi:10.1101/383414.

[4] K. An, et al., Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks, in: Interspeech, 2018, pp. 1913–1917.

[5] J. Wang, et al., Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples, INTERSPEECH, 2016.

[6] B.N. Suhas, et al., Comparison of speech tasks and recording devices for voice based automatic classification of healthy subjects and patients with amyotrophic lateral sclerosis, in: Proc. Interspeech 2019, 2019, pp. 4564–4568, doi:10.21437/Interspeech.2019-1285.

[7] R. Chiaramonte, C. Luciano, I. Chiaramonte, A. Serra, M. Bonfiglio, Multidisciplinary clinical protocol for the diagnosis of bulbar amyotrophic lateral sclerosis, Acta Otorrinolaringologica (English Edition) 70 (2019) 25–31, doi:10.1016/j.otoeng.2017.12.010.

[8] M. Vashkevich, A. Petrovsky, Y. Rushkevich, Bulbar ALS detection based on analysis of voice perturbation and vibrato, in: 2019 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), 2019, pp. 267–272.

[9] A. Tena, F. Claria, F. Solsona, E. Meister, M. Povedano, Detection of bulbar involvement in patients with amyotrophic lateral sclerosis by machine learning voice analysis: diagnostic decision support development study, JMIR Med. Inform. 9 (3) (2021) e21331, doi:10.2196/21331. http://www.ncbi.nlm.nih.gov/pubmed/33688838

[10] A. Tena, F. Clarià, F. Solsona, M. Povedano, Detecting bulbar involvement in patients with amyotrophic lateral sclerosis based on phonatory and time-frequency features, Sensors 22 (3) (2022), doi:10.3390/s22031137. https://www.mdpi.com/1424-8220/22/3/1137

[11] B. Yousefi, J. Shin, E.H. Schumacher, S.D. Keilholz, Quasi-periodic patterns of intrinsic brain activity in individuals and their relationship to global signal, NeuroImage 167 (2018) 297–308, doi:10.1016/j.neuroimage.2017.11.043. https://www.sciencedirect.com/science/article/pii/S1053811917309771

[12] D. Obeid, S. Sadek, G. Zaharia, G.E. Zein, Touch-less heartbeat detection and cardiopulmonary modeling, in: 2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies, 2009, pp. 1–5, doi:10.1109/ISABEL.2009.5373616.

[13] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, 2013, doi:10.1007/978-1-4614-6849-3.

[14] K.N. Batmanghelich, D.H. Ye, K.M. Pohl, B. Taskar, C. Davatzikos, ADNI, Disease classification and prediction via semi-supervised dimensionality reduction, in: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011, pp. 1086–1090, doi:10.1109/ISBI.2011.5872590.

[15] E. Adeli, K.-H. Thung, L. An, G. Wu, F. Shi, T. Wang, D. Shen, Semi-supervised discriminative classification robust to sample-outliers and feature-noises, IEEE Trans. Pattern Anal. Mach.Intell. 41 (2) (2019) 515–522, doi:10.1109/TPAMI.2018.2794470.

[16] A. Tena, ALS models and data repository, (https://github.com/atenad/ALS. Date accessed: November 9, 2021.).

[17] G.L. Pattee, Others, Provisional best practices guidelines for the evaluation of bulbar dysfunction in amyotrophic lateral sclerosis, Muscle Nerve 59 (5) (2019) 531–536, doi:10.1002/mus.26408.

[18] P. Boersma, D. Weenink, Praat: Doing Phonetics by Computer [Computer Program] Version 6.1.01, Technical Report, University of Amsterdam, 2019.

[19] MATLAB and Signal Processing Toolbox Release, The MathWorks, Inc., Natick, Massachusetts, 2021. United States

[20] I. Rodriguez-Lujan, G. Bailador, C. Sánchez Ávila, A. Herrero, G. Vidal, Analysis of pattern recognition and dimensionality reduction techniques for odor biometrics, Knowl.-Based Syst. 52 (2013), doi:10.1016/j.knosys.2013.08.002.

[21] MATLAB, Version 9.9.0.1495850 (R2020b Update 1), The MathWorks Inc., Natick, Massachusetts, 2020.

[22] Max Kuhn, The caret Package, 2009. https://github.com/topepo/caret/.

[23] J.H. Krijthe, RSSL: R package for semi-supervised learning, in: B. Kerautret, M. Colom, P. Monasse (Eds.), Reproducible Research in Pattern Recognition, 2016, pp. 104–115.

[24] Y.-F. Li, Z.-H. Zhou, Towards making unlabeled data never hurt, in: Proceedings of the 28th International Conference on Machine Learning (ICML'11), 2011. Bellevue, Washington

[25] A. Tharwat, Classification assessment methods, Appl. Comput. Inform. (2018), doi:10.1016/j.aci.2018.08.003. http://www.sciencedirect.com/science/article/pii/S2210832718301546