POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering, Biomedical Instrumentation

Master's Degree Thesis

Assessment of Vocal Fatigue of Multiple Sclerosis Patients

Validation of a Contact Microphone-based Device for Long-Term Monitoring

Supervisors

Prof. Alessio CARULLO

Prof. Alberto VALLAN

Candidate

Alice FANTONI

OCTOBER 2023

Abstract

Multiple Sclerosis (MS) is an autoimmune-mediated neuro-degenerative disease of the central nervous system (CNS), involving most motor symptoms such as spasticity, weakness, language disorders and dysphonia; also, fatigue is often considered one of the most debilitating symptom of MS patients. An acoustic analysis of MS is performed with the use of vocal material supplied by the speech therapy and rehabilitation department of Don Gnocchi hospital (Milan). The dataset includes the voice recordings of two balanced subgroups of identical dimension (i.e., sixteen subjects) correspondent to healthy subjects (HS) and MS patients. Speech material consists in the vocalization of the sustained vowel /a/, the reading of a phonetically balanced speech (Notturno) and approximately one minute of free speech for each subject. Additionally, long-term recordings are carried out with VH device only, covering a maximum period of four hours of subjects' daily activities. About the signals acquired with the microphone in air, the phases of pre-processing and harmonic frame selection are executed in Matlab (R2022b) environment and then, the extraction of parameters is operated. With the use of a Logistic Regression (LR) model, data are classified comparing the probability p (returned by the model) to a fixed threshold set to 0.5 and dividing them into two classes (HS and MS). The LR model is trained using a single and a combination of 2, 3, 4 features and the combinations exhibiting the best performance in terms of accuracy and Area Under The Curve (AUC) are selected; then, for these combinations the 5-fold cross-validation is implemented. Best performance in terms of classification are obtained for the reading task (accuracy equal to 92.3%) by selecting 3 features, which are gender, 5° percentile of Cepstral Peak Prominence Smoothed, and harmonic frames ratio V/uV. The expanded uncertainty $U(p)$ of the probability p for each task is evaluated, thus providing a confidence interval; when the confidence interval includes the discrimination probability set to 0.5, the classification of the subject is considered "non-classifiable" and new classification metrics are defined, such as the Realistic Accuracy and the Fraction of Classified (FoC). The implementation of this procedure to the feature combination showing the best performance during cross-validation (gender, $\mathrm{CPPS}_{5,\mathrm{prc}}$ and V/uV), results in FoC of 92.3% and an higher accuracy.

With the aim of validating VH device, considered a valuable aid for long-term monitoring of fatigue, the parameters extracted from the microphone in air are compared to the ones from VH by calculating differences Δ between these measures. Considering sustained vowel /a/ task, the analysis is performed on the parameters local jitter (%), local shimmer (%), $\mathrm{CPPS}_{\mathrm{median}}$ (dB) and $\mathrm{CPPS}_{\mathrm{std}}$ (dB); for balanced and free speech task these differences are carried out for all descriptive statistics of

fundamental frequency f_0 (Hz) and CPPS (dB). For the parameters local jitter, $\text{CPPS}_{\text{median}}$ and CPPS_{std} the validation can be considered passed, while for the others, such as local shimmer, significant differences in terms of Δ values are noted. Since the two devices have different characteristics, i.e. in terms of bandwidth, and they receive as input different signals (the vibration induced by vocal folds for the VH device and the air-pressure signal for the in-air microphone), the use of the VH device requires the definition of specific cut-off values for the extracted parameters. Additionally, the VH device is preferable both for convenience in conducting acquisitions and for its insensitivity to other possible sound sources.
A proposal to assess fatigue is conducted with the use of differences δ between the parameters extracted from the long-term and the short-term monitoring; this comparison is carried out considering the parameters fundamental frequency f_0, CPPS (dB), Background Noise Level (90° percentile) in dBA and Sound Pressure Level (dB). No significant difference in the behavior of the classes with regard to the fatigue experienced is found; however, limitations can be overcome through both an increase in the data-set and in the time interval of the records, being the considered acquisitions too short (between 95 and 200 minutes) to demonstrate fatigue. The parameters acquired with VH for the long-term evaluation are visualized over time and a compensation on the speech intensity value with respect to the noise level is operated.
Eventually, an evaluation of five vocal dose measures as indicators of long-term vocal folds tissue exposure to vibration is provided, which are Time Dose, Cycle Dose, Distance Dose, Energy Dissipation Dose and Radiated Energy Dose. Unless vocal effort is significant among the other subjects, there is the problem of comparing the vocal doses. To compute an assessment at consistent times, a minimum time interval duration common to all subject is considered. Since the limit concerning time interval duration too short for long-term assessment is found, this study is conducted by increasing the time interval and by eliminating the subjects with a short duration.

Acknowledgements

Alla mia famiglia, che mi ha sempre sostenuto in tutto.
Ringrazio le mie amiche Martina e Nicholle per non essere mai lontane.
Grazie a tutti i miei amici e persone vicine
che mi vogliono bene e mi accettano nonostante i difetti.
Infine, ringrazio i miei due compagni di questo viaggio, Geronimo e Beatrice,
che anche se distanti siete stati un punto di riferimento
con cui condividere sfide e successi.

Table of Contents

List of Figures

Chapter 1

Introduction

This chapter provides an overview of the Multiple Sclerosis (MS) disease, which involves a subgroup of the data-set analyzed in this work. In particular, the MS pathology is exposed and then, vocal symptoms and acoustic changes occurring in the affected people are described. The objective in this presentation is to offer a comprehensive view of the main vocal features characterizing the speech material. The physiological systems and mechanisms deputed in the production of human voice are reported in the Appendix A.1.

1.1 Acoustic analysis of MS patients' voice

Multiple Sclerosis (MS) is an autoimmune-mediated neuro-degenerative disease of the central nervous system (CNS), distinguished by inflammatory demyelination with axonal transection. The destruction of myelin sheaths of neurons results in multiple lesions of the brain white matter, brainstem and spinal cord. The diagnosis of MS is based on a combination of clinical findings, imaging and by the demonstration of dissemination of MS disease characteristics in space and time. As a result of lesions throughout the CNS, symptoms involving most motor functional systems may be present. The management of symptoms, such as spasticity, pain, weakness, tremor, cognitive impairment and gait dysfunction, is integral in treatment. Current treatment for MS consist of a multidisciplinary approach, including pharmacological therapies (i.e., disease-modifying therapies (DMTs)), symptomatic treatment, lifestyle modifications, physiological support, and rehabilitation intervention. Since speech is controlled by many areas in the brain, MS lesions can cause several types of voice disorders, ranging from mild difficulties to severe problems. Speech symptoms include alterations such as, language disorders, dysphonia and dysarthria. One pattern that is commonly associated with MS is scanning speech; scanning dysarthria produces speech in which the normal pattern

is disrupted, with abnormally long pauses between words or individual syllable of words. Sometimes, the result of weakness and/or incoordination of the muscles of the tongue, lips, cheeks and mouth led difficulties in being understood and also, speech volume can be affected. Many patients with dysarthria also have dysphagia, which is the difficulty of swallowing; speech therapists and language pathologists are trained to evaluate, diagnose and relieve these problems [1]. Dysphonic symptoms are explained based on the alterations in the neuronal concentration and projection in the periaqueductal gray matter, a finding commonly seen in patients with MS [2]. In order to extrapolate relevant clinical information, an experienced physician may evaluate the phonatory characteristics of an impaired patient; unfortunately, vocal changes in patients with MS are not always perceived even by professional listeners, either due to the intermittency of these changes or on their subtle presence. In this sense, acoustic analysis can assist the physician's perceptual assessment in the early detection of these vocal symptoms or vocal changes. In addition, fatigue is the most common symptom and 40% of patients describe it as the most debilitating one, leading to loss of employment and impairment of activities of daily living. Fatigue can be both, a direct effect of MS or due to secondary causes, such as depression or sleep-related disorders. A proposal to assess this symptom is conducted evaluating long-term acquisitions compared to short-term ones (the last considered as a sort of "baseline"), since fatigue is expected as the recording progresses.

Chapter 2

Materials and methods

This work is carried out in collaboration with the speech therapy and rehabilitation department of Don Gnocchi Hospital in Milan. Don Gnocchi Foundation was established in 1945 under the will of don Carlo Gnocchi and today conducts its activities by relying on the Italian Nation Health Service in twenty-five residential facilities and twenty-seven clinics organized in territorial areas. The mission of the Foundation is to provide for the health and care needs of those in conditions of suffering and fragility, taking care of the patients and the persons called to be there for them, such as family members, health-care professionals and volunteers. The multidisciplinary team of caregivers and health-care professionals of the Foundation provides cures for children with all forms of disabilities, treat with rehabilitation patients of all ages, take care of non-self-sufficient elderly people and terminally ill patients. The speech therapy and rehabilitation department of Don Gnocchi Hospital treats patients with Multiple Sclerosis (MS), which are the class of subjects involved in this thesis work. In particular, the effects of MS disease on voice performance are investigated by analysing a data-set that includes voice recordings of thirty-two subjects: sixteen healthy adults (HS) (mean age 42 years, standard deviation approximately 12 years) and sixteen patients with MS (mean age 44 years, standard deviation approximately 14 years). With the objective of providing a more complete view of the work, the diagram presented in figure 2.1 submits a general outline of all the steps and data treated in different tasks.

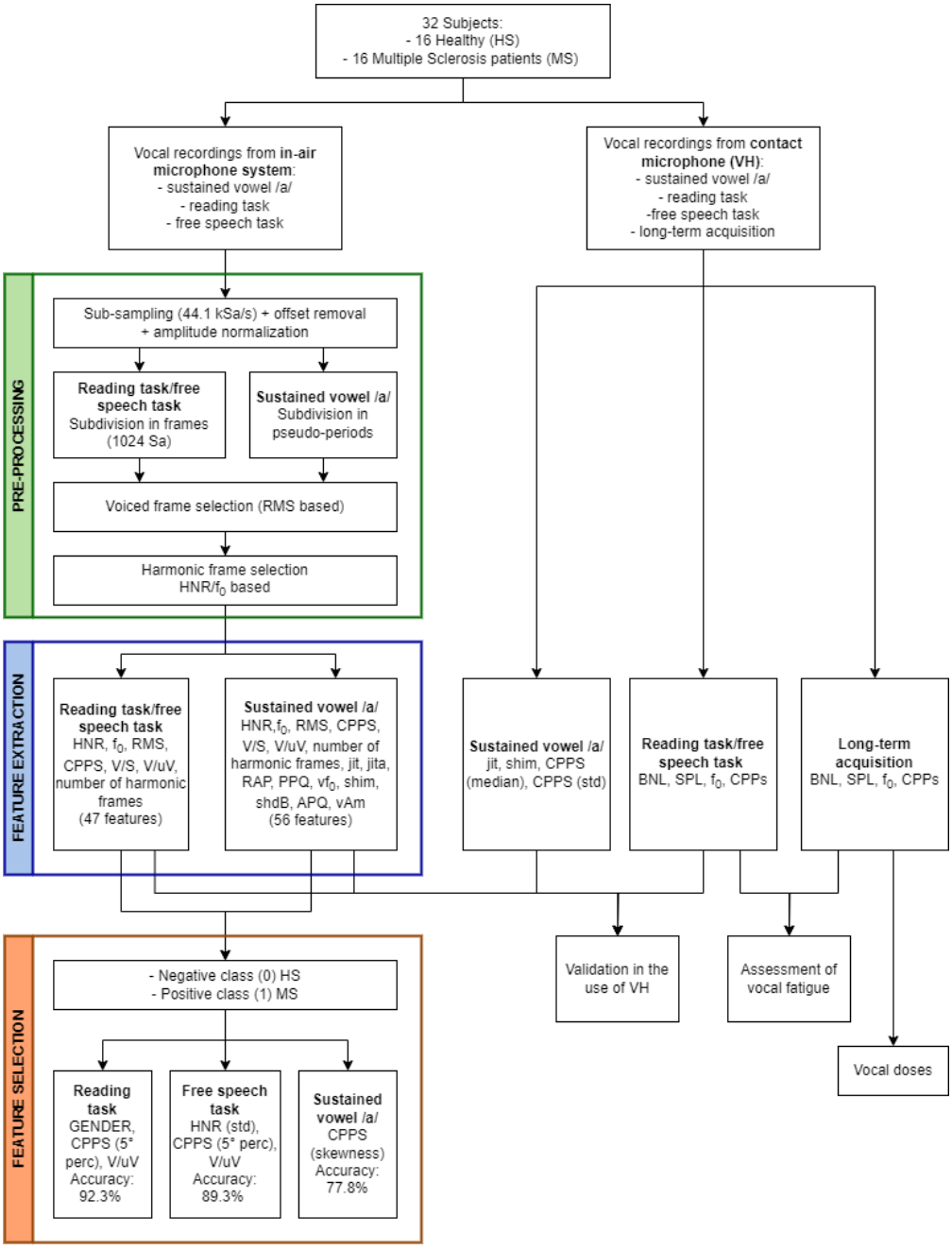

Figure 2.1: Flow-chart showing the various steps performed in the analysis

The voice recordings of the involved subjects are simultaneously performed with a microphone in air and with a contact microphone-based device, hereafter named Vocal Holter (VH). About the speech material, subjects are asked to perform: three

repetitions of the vowel /a/ at a comfortable pitch, level and duration; the reading of a phonetically balanced speech ("Notturno"); an approximately one-minute of free speech. "Notturno", which is reported in Appendix A.2, has the property of being a continuous speech, consisting of phonemes that occurs approximately the same frequency at which they occur in normal conversation in that specific language (that in this case, corresponds to Italian language) [3]. In addition to the mentioned material, long-term monitoring data is also acquired with VH device, comprising the acquisition of a maximum period of 4 hours of subjects' daily activities. About the signals acquired from the microphone in air, the phases of pre-processing and harmonic frame selection are executed in Matlab (R2022b) environment and then, the extraction of parameters is operated. The number of extracted features varies depending on the task under consideration: for the balanced and free speech task 47 features are evaluated, while in the case of sustained vowel /a/ there are 56 available features. Then, with the use of a Logistic Regression (LR) model, data are classified comparing the probability p (returned by the model) of belonging to the positive class to a fixed threshold, which is typically set at 0.5 for binary classifiers. The elements displaying a p higher than the threshold, are assigned to the positive class which corresponds to MS patients, otherwise are attributed to the negative class of HS subjects. The LR model is trained using a single feature and a combination of 2, 3, 4 features and the combinations exhibiting the best performance in terms of accuracy and Area Under The Curve (AUC) are selected; then, for these combinations, the 5-fold cross-validation is implemented with the use of the Classification Learner App in Matlab. In addition, the expanded uncertainty $U(p)$ of the probability p provided by the LR model for each task is evaluated, thus providing a confidence interval for each probability value. When the confidence interval includes the discrimination probability set to 0.5, the classification of the subject is questionable and it is considered "non-classifiable". As a consequence of this decision, new classification metrics are defined, such as the Realistic Accuracy and the Fraction of Classified (FoC).
The main aim of the thesis is to validate the use of VH device as an aid for long-term monitoring of fatigue and vocal dysfunctions. To achieve this purpose, the parameters extracted from the microphone in air are compared to the ones returned from VH by calculating differences Δ between these measures. These comparisons are carried out for the short-term assessment, being the long-term evaluation conducted with the contact microphone-based device only. Considering the three repetitions of sustained vowel /a/, the analysis is performed on the parameters local jitter (%), local shimmer (%), $\text{CPPS}_{\text{median}}$ (dB) and CPPS_{std} (dB); for balanced and free speech task these differences are carried out for all descriptive statistics of fundamental frequency f_0 (Hz) and CPPS (dB). Additionally, a proposal to assess fatigue is conducted with the use again of differences δ between the parameters extracted from the long-term and the correspondent short-term monitoring (i.e.,

balanced and free speech task)), the last considered as a sort of "baseline" (i.e., the parameters during the first instants of the evaluation); this comparison is carried out considering the parameters fundamental frequency f_0, CPPS (dB), Background Noise Level (90° percentile) in dBA and Sound Pressure Level (dB). In this analysis, a difference in the behavior of the two classes with respect to the fatigue is expected. The parameters acquired with VH for the long-term evaluation are visualized over time and a compensation on the speech intensity value with respect to the noise level is operated. The main limitations derive from the number of subjects involved that is low, and in the time time interval of the acquisitions, too short (between 95 and 200 minutes) to compute an assessment on vocal fatigue. Eventually, an evaluation of five vocal dose measures as indicators of long-term vocal folds tissue exposure to vibration is provided, these are the time dose, the cycle dose, the distance dose, the energy dissipation dose and the radiated energy dose. Since the limit concerning time interval duration too short for long-term assessment is found also in the evaluation of the doses, this study is conducted by increasing the time interval and by eliminating the subjects with a short duration.

2.1 In-air microphone system

For the present study, the voices of the involved subjects are recorded using an in-air microphone system placed at a distance from the mouth of 30 cm, characterized with a resolution of 16-bit and setting the sampling rate at 44.1 kSa/s. The in-air microphone records are available in .wav format. The purpose in this first part of the analysis concerns the description of the method used to determine vocal parameters and the identification among these of the most representative ones, leading to the discrimination between the two classes of pathological and healthy subjects.

2.1.1 Pre-processing

The pre-processing phase is performed in parallel for all three tasks (sustained vowel /a/, balanced and free speech). First, all the records are initially listened and analysed with the support of the software Audacity (version 3.2.5), in order to remove the sections at the beginning and at the end of the vocal pieces, which can be considered not relevant for the aim of this work. After that, each signal (which corresponds to one subject) is loaded and re-sampled in Matlab (R2022b) environment at 44.1 kSa/s, which is a value commonly used in literature. Then, the control of the mean value for the entire signal is done. When the mean value is higher than 20% of the root mean square RMS (a.u.) value, it is removed. This step is followed by normalization with the respect to the amplitude: the signal is normalized to the absolute value of the maximum of the analysed signal. After

this stage, a subdivision according to the precise task under analysis is performed: in the case of balanced and free speech, signal samples are grouped into frames of 1024 samples, i.e., a time interval of 23 ms, which is of the same order of magnitude of inter-syllabic pauses. While in the case of the three repetitions of the vowel /a/, signal samples are grouped into pseudo-periods, which are identified through an auto-correlation algorithm. An important operation is implemented removing silence frames and, to enable this, the idea is to use a fixed threshold equal to half of the RMS value of the whole signal. In more detail, frames of 1024 samples are shifted over the signal and if the RMS value of each frame is above the set threshold, then it is considered as a non-silence frame (voiced) and it is saved in an array to perform subsequent processing. Otherwise, if the result is negative then it is a silence frame and can be discarded. The following control is performed selecting harmonic frames according to a specific criterion. This criterion consists in extracting the parameters HNR and f_0 value (the description of the determination of these two measures will be exposed later in this chapter) of each voiced frame and selecting only the frames that exhibit a value of HNR greater than 0 dB and that have a frequency jump between adjacent frames not lower than -25% and not higher than +50%. Thanks to this check, only frames that are characterized by harmonic content not lower than the noise energy in each frame will be promoted to the next part of this study, that corresponds to the feature extraction step.

2.1.2 Feature-extraction

The extraction of the parameters coming from the spectral, cepstral and time domain, is performed for each block signal selected during the pre-processing phase. As already stated in the previous section, two different operations are used for the three cases under analysis: in the case of balanced and free speech, the signal is evaluated with a window length of 1024 samples, while in the case of sustained vowel /a/ is preferable to use pseudo-periods as frames. Only the parameters related to the harmonic frames of the whole signal, are saved and considered to implement feature extraction. In this way, the collected sequences of the extracted parameters are transformed into a statistical distribution, which can be described with statistical metrics, in order to reduce the size of the collected data and, to achieve a more representative observation. In particular, the nine descriptive statistics of mean, median, mode, 5-th percentile, 95-th percentile (as indices of central tendency), range, standard deviation (as measures of variability) and skewness and kurtosis (as shape parameters) are calculated. For all three tasks, the distribution of four acoustic parameters (i.e., Harmonic to Noise Ratio, fundamental frequency, Root Mean Square and Cepstral Peak Prominence Smoothed), represented with the already mentioned statistics are evaluated; only in the case of the sustained vowel /a/ task, nine perturbation parameters are considered. Furthermore, three

extracted parameters have been added for all the tasks. The figures 2.2 and 2.3 summarize the extracted parameters according to the task under observation. The description of the perturbation parameters and their implementation in Matlab scripts refers to the software instruction manual of MDVP, Model 5105 [4]. For the other recorded metrics and the acoustic parameters, their definition is presented in the section that follows.

BALANCED/FREE SPEECH TASK
Harmonic to Noise Ratio, HNR (dB)
Fundamental frequency, f_0 (Hz)
Root Mean Square, RMS (a.u.)
Smoothed Cepstral Peak Prominence, CPPS (dB)
Non-silent frames ratio, V/S (%)
Harmonic frames ratio, V/uV (%)
Number of harmonic frames

Figure 2.2: Extracted features for balanced and free speech task

SUSTAINED VOWEL /a/
Harmonic to Noise Ratio, HNR (dB)
Fundamental frequency, f_0 (Hz)
Root Mean Square, RMS (a.u.)
Smoothed Cepstral Peak Prominence, CPPS (dB)
Non-silent frames ratio, V/S
Harmonic frames ratio, V/uV
Number of harmonic frames
Absolute Jitter, Jita (μs)
Jitter Percent, Jitt (%)
Relative Average Perturbation, RAP (%)
Pitch Period Perturbation Quotient, PPQ (%)
Coefficient of Fundamental Frequency Variation, vf_0 (%)
Shimmer , ShdB (dB)
Shimmer Percent, Shim (%)
Amplitude Perturbation Quotient, APQ (%)
Coefficient of Amplitude Variation, vAm (%)

Figure 2.3: Extracted features for sustained vowel /a/ task

Acoustic parameters

Harmonic to Noise Ratio HNR and fundamental frequency f_0

Harmonic to noise ratio is the average ratio of the harmonic spectral energy and it is evaluated in dB. The determination of the HNR and the fundamental frequency f_0 is computed with the method of the auto-correlation (AC) [5]. Considering a stationary (i.e., its statistics are constant) time signal $x(t)$, the auto-correlation function of the lag τ is defined as:

$$AC = \int x(t)x(t+\tau)\,dt \tag{2.1}$$

This function has a global maximum at zero lag, which corresponds to the power of the signal. To evaluate the harmoniousness of the voice signal, the parameter HNR is calculated as:

$$HNR = 10\log_{10}\frac{AC(T)}{AC(0) - AC(T)} \tag{2.2}$$

where $AC(T)$ is the auto-correlation function at lag T and $AC(0)$ is the auto-correlation function at zero lag. Performing a normalization of AC of the signal to the maximum value of AC at zero lag, follows that, at the numerator the relative power of harmonic components is expressed, while at the denominator the relative power of noise is derived. The normalized HNR formula can be rewritten as:

$$HNR = 10\log_{10}\frac{AC(T)/AC(0)}{1 - AC(T)/AC(0)} \tag{2.3}$$

The evaluation of HNR is computed on non-silent frames only, to discriminate valid (harmonic) from non-valid (unharmonic) signal frames. The fundamental frequency f_0 of a periodic signal of period T, is defined as $f_0 = 1/T$. While for healthy voices, the HNR value is in most of the cases above 0 dB, pathological voices can sometimes lead to negative values of HNR, i.e., the energy of the harmonic component is lower than the noise level. In this work, only frames that exhibited an HNR value higher than 0 dB were selected. As previously mentioned, an additional condition on frequency jumps is set, considering frames as invalid if they differ more than half an octave. The fundamental frequency f_0 of a periodic signal of period T, is defined as $f_0 = 1/T$. Regarding the voice signal, f_0 represents the number of cycles produced by the vocal folds per second. Being the vocal folds of men and women different both in size and vibration, this results in different phonation (see the Appendix A.1 for more details). Several studies report that male's vocal tract is longer than female's, and their vocal folds are thicker and larger; consequently, they vibrate at approximately one-half the frequency of women's during phonation, thus producing a lower fundamental frequency [6] [7]. The *gender* in this work is

considered along with the other extracted parameters as a feature for each signal under analysis in all three tasks, and a sub-division between frequency ranges is performed: from 75 Hz to 300 Hz for males and from 100 Hz to 400 Hz for females.

Root Mean Square RMS

The RMS value x_{RMS} of a set of n signal samples $\{x_1, x_2, ..., x_n\}$ is defined as the square root of the arithmetic mean of the squares of the samples:

$$x_{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{2.4}$$

Being the operation of square root, computed only after the normalisation in amplitude of the signal, the RMS parameter is expressed in "arbitrary units" (a.u.). The RMS value is employed in the pre-processing phase to discriminate the silence frames. As previously mentioned, silence frames are removed from the signal throughout the use of a threshold, equal to half of the RMS value of the entire signal. A 1024-samples window is shifted over the recorded signal, to verify if the RMS value of each window is above the threshold. If the result is negative, then it is considered a silent frame and it is removed, otherwise voiced frames are saved in an array.

Cepstral Peak Prominence Smoothed CPPS

To obtain a complete overview of the parameter, it is first important to expose what the Cepstrum is and how it is obtained. The term derives from the anagram of the word spectrum, and mathematically is the spectrum of a logarithmic spectrum of a time waveform. With more detail, to produce a cepstrum, a Fourier transformation of an acoustic signal is performed first to create a spectrum, and the voice signal is transformed from the time domain to the frequency domain. The first power spectrum shows the frequency distribution of the signal energy. Subsequently, performing a Fourier transformation of the spectrum, produces the cepstrum. In doing so, the signal is transformed from the frequency domain to the quefrency domain (which is equal to 1/frequency), and the second spectrum offers a better understanding of how periodic the harmonic components are [8] [9]. The Cepstrum is represented in Eq. (2.5):

$$\mathcal{C}_p(\tau) = |\mathcal{F}\{\log(|\mathcal{F}\{x(t)\}|^2)\}|^2 \tag{2.5}$$

where $x(t)$ is the voice signal, $\mathcal{F}$ is the Fourier transformation, $|\mathcal{F}\{x(t)\}|^2$ is the signal power spectrum and, τ is the quefrency (that is the anagram of the word frequency) and it is a measurement of time, but in the cepstrum domain. The

rhamonics (i.e., the cepstrum peaks in the domain of quefrency) occur at the quefrency at which the original time waveform has the fundamental frequency. While the parameters in the time domain (e.g., jitter and shimmer, presented later in this chapter), whose limitation is to depend on the accurate identification of cycle boundaries (i.e. where a cycle of vocal folds vibration starts and finishes) and thus, they become unreliable with highly perturbed signals, by switching the signal in the quefrency domain attempts to circumvent this problem [10]. Taking a step back on the production of the human voice, the speech signal $x(t)$ can be defined as the convolution in time of the following components; $g(t)$ that is related to the glottal pulses (modelled as a long train of glottal pulses), $v(t)$ that represents the impulse response of the vocal tract and lastly, $r(t)$ which is associated to the effect of the acoustic wave radiation at the lips (modelled as an impulse response):

$$x(t) = g(t) * v(t) * r(t) \tag{2.6}$$

Performing the Fourier transformation of the quasi-periodic signal $x(t)$ and therefore, by switching to the frequency domain, the operation of convolution becomes a product. Then, the power spectrum can be obtained:

$$X(f)^2 = G(f)^2 \cdot H(f)^2 \tag{2.7}$$

where $H(f)$ represents the combined result of vocal tract and lip radiation [11]. The logarithm, which can be easily isolated through a filter, allows to convert the product of the Eq. (2.7) into a sum and thus, become an aid in separating the two components of the speech signal. Hence, the concept of liftering (i.e., linear filtering) of the log spectrum, as a way of emphasizing the periodic component of the log spectrum, is applied to enhance the ability to detect echoes from a signal. The cepstrum can be used in the analysis of voice signal for pitch detection, in fact the quefrency at which the cepstral peak occurs, provides information on the fundamental frequency of the considered frame, whereas, the prominence with respect to the level of "background" noise indicates the harmoniousness of the signal. In addition to this, the cepstrum is also useful for obtaining information on the spectral envelope of the signal for speech analysis. From cepstrum two important parameters to assess the quality of voice can be defined, namely the cepstral peak prominence (CPP) and, its smoothed version (CPPS). The CPP is the difference in amplitude between the cepstral peak and the corresponding value on the linear regression line directly below the peak. CPP is, thus, a measure of the degree of harmonic organization. An healthy voice, which has a well-defined harmonic structure, has a strong cepstral peak, while signals lacking a well-defined harmonic structure (i.e., in the case of a pathological voice) have small CPPS. CPPS considers two smoothing steps before calculating the cepstral peak prominence. Both parameters are expressed in dB and they show to correlate with perception

of breathiness, with CPPS being the best predictor. Being both CPP and CPPS based on a peak-to-average calculation of the fundamental frequency and, not relying on the accurate determination of it, tend to be more reliable than other measures of periodicity. For the implementation of the CPPS algorithm, a Matlab (R2022b) script is developed. Signals are sampled at a frequency of 22050 Hz and CPPS is computed every 2 ms frame length, using a 1024-point Hanning-type analysis window (of duration 46 ms). For each window, a series of steps lead first to the cepstrum domain and, then, to the peak prominence estimation. Starting from the signal in the time domain, the Fast Fourier Transform (FFT) algorithm is computed on the windowed signal, in order to obtain the spectrum amplitude and lastly, the FFT algorithm is performed again on the log power spectrum to reach the cepstrum domain. Furthermore, two smoothing steps are implemented on the obtained cepstra: the cepstra of each considered window are smoothed in time using a 14 ms (which corresponds to seven frames); then, the smoothing in quefrency is performed using a 7-bin averaging window. In the following step, the regression line is calculated between 1 ms and the maximum quefrency value. The first millisecond is excluded, since a property of the cepstrum is that at low quefrencies is more affected by the spectral slope, than by the spectrum periodicity. As already mentioned, the CPPS is evaluated as the difference (in dB) between the peak in the cepstrum and the corresponding value at the same quefrency on the regression line. Since the quefrency at the cepstral peak generally corresponds to the inverse of the fundamental frequency, the cepstral peak is searched between 3.3 ms (which corresponds to 300 Hz) and 16.7 ms (i.e., 60 Hz); this is performed in order to include the typical fundamental frequency range of female and male adults [12].

Other recorded metrics

In a first part of the work, the signal is analysed for the detection of voiced and unvoiced frames. As mentioned before, silent and non-silent frames are identified with the use of the RMS value, in this case the number of silent frames is taken into account, and finally, they are removed. The second step is the selection of harmonic and non-harmonic frames: this is achieved by verifying the HNR value of the considered frame and then, checking frequency jumps. Based on this analysis, it is possible to obtain three other metrics that will be considered as classification parameters:

- Non-silent frame ratio (%):

$$V/S = 100 \cdot \frac{n_{voiced}}{n_{voiced} + n_{unvoiced}} \tag{2.8}$$

  where the number of voiced frames (n_{voiced}) includes both harmonic and non-harmonic frames, and ($n_{unvoiced}$) indicates the number of silent frames.

- Harmonic frame ratio (%):

$$V/uV = 100 \cdot \frac{n_{harmonic}}{n_{harmonic} + n_{non-harmonic}} \tag{2.9}$$

- Length: the number of harmonic (valid) frames after the pre-processing phase.

Period and amplitude parameters

In contrast to balanced and free speech tasks, in the case of the repetitions of three /a/ phonemes, nine stability parameters in period and amplitude are also derived. These parameters provide an assessment of the stability in period and in amplitude of the vocal signals by measuring the variations in these quantities from cycle-to-cycle.

Absolute Jitter Jita

It is an absolute measure in microseconds (μs) of the period-to-period variability of the pitch period with the exclusion of voice breaks.

$$Jita = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}| \tag{2.10}$$

where $T_0^{(i)}$, $i = 1,2,...N$ are the extracted pitch period data and N the number of extracted pitch periods.

Jitter Percent Jitt

It is a relative evaluation of the period-to-period variability of the pitch.

$$Jitt = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}|}{\frac{1}{N} \sum_{i=1}^{N} T_0^{(i)}} \tag{2.11}$$

where $T_0^{(i)}$, $i = 1,2,...N$ are the extracted pitch period data and N the number of extracted pitch periods. Both Jita and Jitt represent the measurements of the same type of pitch perturbation. While Jita is an absolute measure, which makes it strongly related to the fundamental frequency of the voice signal (i.e. higher pitch results into lower Jita, thus, normative values for males and females differ), in the case of Jitt this dependency is significantly reduced.

Relative Average Perturbation RAP

It is a relative measurement of the irregularity of the pitch period of the voice with a smoothing factor of 3 periods.

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}|\frac{T_0^{(i-1)}+T_0^{(i)}+T_0^{(i+1)}}{3} - T_0^{(i)}|}{\frac{1}{N}\sum_{i=1}^{N}T_0^{(i)}} \tag{2.12}$$

where $T_0^{(i)}$, $i = 1,2,...N$ are the extracted pitch period data and N the number of extracted pitch periods.

Pitch Period Perturbation Quotient PPQ

It is a relative evaluation of the period-to-period variability of the pitch within the signal with a smoothing factor of 5 periods.

$$PPQ = \frac{\frac{1}{N-4}\sum_{i=1}^{N-4}|\frac{1}{5}\sum_{r=0}^{4}T_0^{(i+r)} - T_0^{(i+2)}|}{\frac{1}{N}\sum_{i=1}^{N}T_0^{(i)}} \tag{2.13}$$

where $T_0^{(i)}$, $i = 1,2,...N$ are the extracted pitch period data and N the number of extracted pitch periods. Jita, Jitt, RAP, PPQ are parameters used to quantify the frequency variation from cycle-to-cycle within the vocal signal (stability in frequency). Cycle-to-cycle irregularity can be related with the inability of the vocal cords to support a periodic vibration with a defined period, typical of hoarse and breathy voices.

Coefficient of Fundamental Frequency Variation vf_0

It represents the relative standard deviation of the fundamental frequency and reflects any variation of f_0 within the voice signal.

$$vf_0 = \frac{\sigma}{f_0} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N}\sum_{j=1}^{N}f_0^{(j)} - f_0^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N}f_0^{(i)}} \tag{2.14}$$

where $f_0 = \frac{1}{N}\sum_{i=1}^{N}f_0^{(i)}$ and $f_0^{(i)} = \frac{1}{T_0^{(i)}}$ are the period-to-period fundamental frequency values, $T_0^{(i)}$, $i = 1,2,...N$ are the extracted pitch period data and N the number of extracted pitch periods.

Shimmer ShdB

It measures in dB the very short term (cycle-to-cycle) irregularity of the peak-to-peak amplitude of the voice.

$$ShdB = \frac{1}{N-1} \sum_{i=1}^{N-1} |20 \log(\frac{A^{(i+1)}}{A^{(i)}})| \tag{2.15}$$

where $A(i)$, $i = 1,2,...N$ are the peak-to-peak amplitude data and N the number of extracted impulses.

Shimmer Percent Shim

It is a relative measure of the period-to-period variability of the peak-to-peak amplitude within the voice signal.

$$Shim = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A^{(i)} - A^{(i+1)}|}{\frac{1}{N} \sum_{i=1}^{N} A^{(i)}} \tag{2.16}$$

where $A(i)$, $i = 1,2,...N$ are the peak-to-peak amplitude data and N the number of extracted impulses. Although Shim and ShdB use different measures for the result (i.e., percent and dB), both are relative evaluations of the same type of amplitude perturbation.

Amplitude Perturbation Quotient APQ

It is a relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analysed voice sample at a smoothing of 11 periods.

$$APQ = \frac{\frac{1}{N-10} \sum_{i=1}^{N-10} |\frac{1}{11} \sum_{r=0}^{10} A^{(i+r)} - A^{(i+5)}|}{\frac{1}{N} \sum_{i=1}^{N} A^{(i)}} \tag{2.17}$$

where $A(i)$, $i = 1,2,...N$ are the peak-to-peak amplitude data and N the number of extracted impulses. ShdB, Shim and APQ are measures of shimmer by showing the irregularity of the peak-to-peak amplitude of the voice.

Coefficient of Amplitude Variation vAm

It reveals any variations in the cycle-to-cycle amplitude of the voice.

$$vAm = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^{N} (\frac{1}{N} \sum_{j=1}^{N} A^{(j)} - A^{(i)})^2}}{\frac{1}{N} \sum_{i=1}^{N} A^{(i)}} \tag{2.18}$$

where $A(i)$, $i = 1,2,...N$ are the peak-to-peak amplitude data and N the number of extracted impulses. Either random or regular short-term or long-term variations increase the value of vAm.

2.1.3 Classification-based feature selection

The feature selection (FS) process is based on the evaluation of the performance of a Logistic Regression (LR) model (described in Appendix A.3), which is trained in Matlab R2022b environment, using a different number of input features. The number of features employed, varies depending on the task under consideration: while in the case of balanced and free speech task there are 47 available features, for the three repetitions of the vowel /a/ the number of features become 56 (see section 2.1.2 for more details). However, it is valuable to report that only the ones with low correlation and deemed to be statistically significant are considered [13]. The LR algorithm implemented in this study uses the Matlab built-in function *fitglm*, which receives as input the data to be classified, the real class (0 or 1) and other model specifications to define the distribution of the response variable as binomial ('Distribution', 'binomial') and to set the logit function (shown in Eq. (A.1)) as the link function ('Link', 'logit'). The function returns the LR model, the probabilities and an evaluation of the coefficient standard errors and covariances, useful to estimate the regression model's performance. The algorithm first selects the number of k features to be combined and, subsequently, creates the combinations through the binomial coefficients of the Eq. (2.19). About the features combinations, in this step of the work, the Matlab (R2022b) function *nchoosek* is used, which returns both the binomial coefficients and all the combinations. The function receives as input the number k of features considered (which can be 1, 2, 3 and 4) and the total number of features n (previously mentioned as 47 for the balanced and free speech task and 56 for the vowel /a/). A summary of the number of non-repeated feature combinations obtained is presented in figure 2.4 for all three tasks. As shown, when the number of considered features k rises, also the computational cost of the algorithm in the Matlab environment becomes higher, which is why the number of features chosen is set to range from 1 to 4.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.19}$$

Before proceeding in the training of the LR model with a single feature or a combination of 2, 3 or 4 features, a check on the ρ^2 values of each combination of features and their p-values is computed. The coefficient of determination ρ^2 is calculated for each feature as the square value of the correlation, performed with the use of the Matlab function *corr* which also returns the p-values of the correlations. First, if the value of correlation coefficient is between 0 and 0.5 and the corresponding p-value is lower than 0.05, then the combination of features is accepted. This first check is done for all possible combinations of features and if the condition is satisfied, then the pair, triplet or quadruplet is considered valid, and it is sent as input to the LR model. When the model is obtained, a second

check is performed on the p-value of the identified parameters β_i: in this case a fixed threshold higher than the previous one is used, since it is noted that with a p-value equal to 0.05 no combination of features is selected [14]. The model returns for each observation a probability value between 0 and 1, which with the use of a fixed threshold set at 0.5, is converted into a binary value 0/1 which represents the membership (predicted by the model) of an element to one of the two classes. The elements characterized by p higher than the threshold (i.e., $p > 0.5$) are assigned to the positive class (Class 1), otherwise if the probability is lower, to the negative class (Class 0). In this work, the negative class is associated with the healthy control group (HS) and the positive Class 1 with the group of patients with Multiple Sclerosis (MS). To obtain the *confusion matrix* (see the Appendix A.3 for more detail), the probabilities returned by the built-in function *predict* are compared to the real responses. Also, with the confusion matrix the typical metrics of classification performance for each combination are derived, these are Accuracy, Specificity, Sensitivity, Precision and AUC (their definition is reported in the Appendix A.3). Lastly, at the end of the FS algorithm a combination of features (for each k) with the best performances (which will be used later during the validation phase) are selected; to allow this, the combination of features with the highest accuracy value is chosen and in the case there are more than one with the same maximum accuracy value, the one with the highest AUC value is taken, being the latter a good indicator of the prediction capability of each trained model.

SELECTED FEATURES	BALANCED/FREE SPEECH	SUSTAINED VOWEL /a/
k=1	47	56
k=2	1081	1540
k=3	16215	27720
k=4	178365	3672900000

Figure 2.4: Number of combinations for all three tasks

2.1.4 Model validation

The validation of the logistic regression model is performed after FS process with the use of the Classification Learner App, within the Matlab (R2022b) environment. k-fold cross validation with $k = 5$ is implemented in order to avoid overfitting errors and it is most often used when the data-set is not very large in size. This algorithm consists of partitioning the data into five subsets of equal size, using one subset at a time in rotation to validate the model and the other four subsets to train the model; being repeated five time, at the end all five subsets are used once as validation set. The steps performed during the validation phase consist of,

first, loading the input data matrix containing the features and their class in the Classification Learner App: the matrix of healthy subjects HS and the matrix of patients MS are combined together, and each subject HS/MS has assigned a label 0/1 that corresponds to their class of membership; after that, the combination of features with the highest performance previously selected, is manually inserted in the interface and the validation of the LR model is performed. The App allows to choose different classification models from several available, including Logistic Regression. The Classification Learner App returns as output the confusion matrix relative to the validation phase of the model and within this information it is possible to obtain the values of all the classification metrics, since the App provides only Accuracy, ROC curve and the relative AUC value. In addition, there is the possibility for each feature combination to recreate the validated model in Matlab environment with the command "Generate Function" in the Export section.

2.1.5 Expanded uncertainty analysis of the LR model

Eventually, the final step of the work regarding the vocal analysis from in-air microphone system, is the evaluation of the expanded uncertainty $U(p)$ of the probability p returned by the LR model in the Classification Learner App in Matlab (R2022b). The performance metrics described in section 2.1.3 were obtained comparing the probability expressed by Eq. (A.2) to the default threshold equal to 0.5. However, the standard uncertainties of the regression coefficients β_i and their mutual covariances can be taken into account to provide a more realistic performance evaluation. With this method, the variance $u^2(p)$ can be estimated according to the approach proposed in [15]:

$$u^2(p) = \sum_{i=1}^{N} (\frac{\partial p}{\partial \beta_i})^2 \cdot u^2(\beta_i) + 2 \cdot \sum_{j=1}^{N-1} \cdot \sum_{k=j+1}^{N} (\frac{\partial p}{\partial \beta_j}) \cdot (\frac{\partial p}{\partial \beta_k}) \cdot u(\beta_j, \beta_k) \tag{2.20}$$

where $u(\beta_i)$ is the standard uncertainty of each regression coefficient, $u(\beta_j, \beta_k)$ is the covariance of each couple of coefficients β_j and β_k. The sensitivity coefficients of p are obtained by computing partial derivatives with respect to the regression coefficients β_i as expressed in:

$$\begin{aligned} (\frac{\partial p}{\partial \beta_0}) &= p \cdot (1-p) \\ (\frac{\partial p}{\partial \beta_i}) &= F_i \cdot p \cdot (1-p);\ i \geq 1 \end{aligned} \tag{2.21}$$

where F_i represent the selected features [16] [14]. The expanded uncertainty $U(p)$ is obtained by multiplying the positive square root of Eq. (2.20) by a coverage factor of 2; this allows the creation of confidence intervals $[p - U(p) \div p + U(p)]$ for each

probability value returned by the LR validated model. The expanded uncertainty is graphically represented by vertical error bars as shown in figure 2.5. Observing this

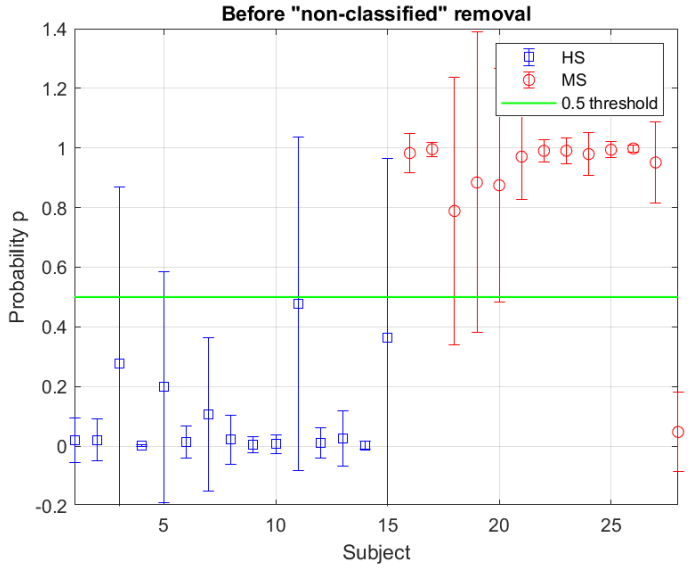

Figure 2.5: Example of confidence intervals for both healthy and pathological classes.

example, there are subjects whose confidence interval includes the discrimination probability threshold set at 0.5, thus making questionable their involvement in the binary classification. In these cases, the decision is made to tag these subjects as "non-classifiable" and therefore, a third class is introduced. Subsequently, in order to have an objective evaluation of the effect of "non-classified" subjects on previously obtained classification performance, new metrics are introduced such as the Realistic Accuracy ($Accuracy_R$) and the Fraction of Classified (FoC). The Realistic Accuracy is nothing different but the calculation of the accuracy as performed in Eq. (A.3) by excluding the elements belonging to the third class of "non-classified". A further parameter defined to evaluate the significance of realistic performance, is the Fraction of Classified which is the faction of classified elements with respect to the total number of elements.

2.2 Vocal Holter (VH) device

The Vocal Holter (VH) device is a monitoring system that allows a relationship between daily voice use and voice disorders to be assessed. The VH kit is composed by three elements, which are: the Data Acquisition and Processing (DAP) unit, which embeds a microphone in air and a spacer in order to maintain the subject's mouth at a fixed and known distance from in-air microphone during calibration; the contact microphone (model hx-505-1-1); the power adapter with its cable. The contact microphone measures the skin vibrations caused by the vocal folds' activity [17]. It is important to place the device around the subject's neck, making sure that the ending parts of the collar adhere as much as possible to the skin area above vocal cords; moreover, once the device is placed, it should be comfortable for the user because it can not be moved for the entire duration of the recording. Since each subject has different body characteristics, VH device is also equipped with a pin allowing to widen or tighten the collar depending on the size of the neck. The device samples the signal induced by vocal folds' activity at a rate of 44.1 kSa/s using 16 bit of resolution. It is noted that, with respect to the microphone in air, the use of a contact microphone-based device allows to minimize the effects of sound sources different from the signal of interest. The samples acquired with VH are grouped into frames of approximately 46 ms and only voiced frames are processed [18]. In addition, it is reported that VH device uses the same method of the microphone in air for the harmonic frames selection (presented in section 2.1.1). In order to correctly use the VH device, the subsequent steps are followed: connect the contact microphone to the DAP unit and make sure the subject wears the collar arund the neck; turn on the DAP unit and connect the PC to Wi-Fi access point created by the device; open the web interface on the PC and select the operation to be performed; once the vocalization is completed, the DAP unit returns the data and as a result, a message window is shown with the value of the estimated parameters. Data stored in the internal memory of the DAP unit are rendered available in .txt format. Either short-term or long-term vocal quality assessment can be performed with the VH device. Regarding short-term assessment, this consists in the same vocal material that is used with the microphone in air; three repetitions of the vowel /a/ at a comfortable pitch, level and duration; the reading of a phonetically balanced speech ("Notturno"); an approximately one-minute free speech. The VH device can measure vocal and environmental parameters; regarding the vocal parameters, these are the Sound Pressure Level (SPL) in dB, the fundamental frequency f_0 in Hz, the Voicing Time Percentage PPT (%), which is defined as the percentage of time spent phonating during the total monitoring period, the Smoothed Cepstral Peak Prominence in dB and, the local jitter and shimmer both expressed in % (the last two concerning only the uttering of a continuous vowel). In order to estimate the speech SPL of the speaker at a fixed distance d_0 of 22 cm in front of

the mouth, each subject has to perform a preliminary calibration, which consists in the repetition of the vowel /a/ at increasing levels in front of the in-air microphone integrated in the DAP unit. As environmental parameters the Background Noise Level (BNL) in dB, the air temperature (θ) in °C and the air relative humidity, are derived. All these parameters vary depending on the task under analysis or on the type of assessment undergoing (short-term/long-term evaluation). With respect to in-clinic recording of voice, in-field long-term monitoring can offer insight into fatigue assessment and can help to identify vocal dysfunctions; in this work, with the use VH device long-term vocal quality evaluation is considered, comprising the records of a maximum period of 4 hours of subjects' daily activities. After the subject vocally performs either short-term or long-term assessment, the data are processed by the DAP unit of VH and can be downloaded with the implementation of a proper script in Matlab (R2022b) environment. In addition, the same data-set of subjects already used for the in-air microphone evaluation, is employed with VH device. The main aim of the investigation is to validate the use of VH as a tool, which on the same level as the in-air microphone, is useful in providing distributional parameters that are able to characterize vocal health. This objective is carried out through the calculation of differences Δ between the parameters extracted from the in-air microphone and the device VH.

Another study that is undertaken is to measure the fatigue reported by subjects during long-term monitoring with respect to short-term (i.e., balanced and free speech task); this comparison is carried out with the use of differences δ, between the parameters extracted from the two types of assessments. In addition to this, voice parameters over time in relation to the environment are visualized for each subject and finally the differences between the two classes of HS and MS are considered. Eventually, efforts are focused on the evaluation of vocal dose measures as indicators of the long-time exposure of the vocal folds tissue to vibrations (i.e., vocal load).

2.2.1 Comparison of parameters extracted from the microphone in air

In this part of the work all the various steps aimed at validating the use of VH for monitoring voice quality are presented. The concept behind this evaluation is that, if acquisitions does not only take place during in-clinic consultation, but are performed during subjects' daily activities, it is important to use a device that can exclude the background noise, which normally occurs in an uncontrolled environment; a contact microphone-based device can accomplish this task. Based on this assumption, the question arises whether VH provides comparable (or at least congruent) information to the one obtained from the microphone in air. With this purpose, the analysis is performed by calculating the differences, indicated with

greek capital letter Δ, between the extracted parameters (obtained in section 2.1.2) from the microphone in air (which will be simply called in this study as MIC in order to differentiate it) and the parameters that are stored inside the DAP unit acquired with VH. To better clarify this concept, the figures presented in this section derive from the application of the Eq. (2.22) to all the parameters available for a certain task:

$$\Delta = MIC - VH \tag{2.22}$$

These comparisons are carried out for the short-term assessment (being the long-term evaluation conducted with VH only), which includes the three repetitions of sustained vowel /a/, the balanced and free speech task.
Considering the three repetitions of sustained vowel /a/, the parameters available to operate this comparison are: local jitter (%), local shimmer (%), $\text{CPPS}_{\text{median}}$ (dB) and CPPS_{std} (dB) (listed in figure 2.6). In the case of balanced and free speech task these differences are carried out for all descriptive statistics of fundamental frequency f_0 (Hz) and CPPS (dB) (as shown in figure 2.7). It is valuable to highlight the fact that, while in the case of sustained vowel /a/ only deltas relative to median and standard deviation of the parameter CPPS are calculated, considering reading and free speech task all the nine descriptive statistics of the extracted parameters are involved (i.e., mean, median, mode, 5-th percentile, 95-th percentile, range, standard deviation, skewness and kurtosis). This difference is indicated with a red asterisk * in the figure relative to the sustained vowel /a/ task.

SUSTAINED VOWEL /a/
Jitter Percent, Jitt (%)
Shimmer Percent, Shim (%)
Smoothed Cepstral Peak Prominence, CPPS (dB) (median)*
Smoothed Cepstral Peak Prominence, CPPS (dB) (std)*

Figure 2.6: Parameters used to perform the comparison between the two devices for sustained vowel /a/ task

BALANCED/FREE SPEECH TASK
Fundamental frequency, f_0 (Hz)
Smoothed Cepstral Peak Prominence, CPPS (dB)

Figure 2.7: Parameters used to perform the comparison between the two devices in the case of balanced and free speech task

Sustained vowel /a/

Being considered only the parameters summarized in figure 2.6, which are in common with the two devices, their number is low. Generally, the following steps, carried out to obtain deltas, are the same for all the short-term evaluation tasks. The matrix relative to the parameters extracted with the in-air microphone and the one relative to the contact microphone-based device are loaded, these four matrices (two corresponding to healthy subjects and two relative to patients with MS) contain the mean value of the three repetitions for each subject. To provide a better coordination between the two classes, the matrices are concatenated resulting in one matrix for all the subjects relative to MIC and one matrix relative to VH. Then for each parameter, deltas are determined, and also their mean and standard deviation are considered. Deltas for each subject are graphically represented (see an example in figure 2.8), where red elements are related to MS class and the blue ones correspond to HS class. In order to have more tangibility of the dispersion of the values of deltas, the standard deviation of the mean value, or standard error (indicated as σ), is obtained as a ratio between the standard deviation and the root square of the total number of subjects in the experiment. Also, the confidence interval chosen for each parameter is set to $\pm 1^{*}\sigma$, which is more significant, since the number of subjects involved is low (i.e., the variability is reduced if the number of subjects is higher). The correspondent values associated with the black line (representing the mean value of delta of the considered parameter) and with the green lines (which define the confidence interval) are reported in legend (in figure 2.8). Deltas provide information on how much the parameters extracted from the in-air microphone differ from those processed by the contact microphone-based device (i.e., the variability in the estimation of the parameters between the two devices).

Balanced and free speech task

Regarding the balanced and free speech task, the various steps performed to achieve delta parameters are the same already presented in the case of the sustained vowel /a/ task. Also in this case, the two matrices correspondent to the parameters extracted after the pre-processing phase for the microphone in air and the parameters provided by the VH device are loaded (figure 2.7). Subsequently, the metrics in common with the two devices are selected, and after having obtained two matrices relative to MIC and two relative to VH, differences (and their mean value and standard deviation) are calculated for each parameter. What previously said for the legend description in the case of sustained vowel /a/ task, also applies to deltas considered in this investigation.

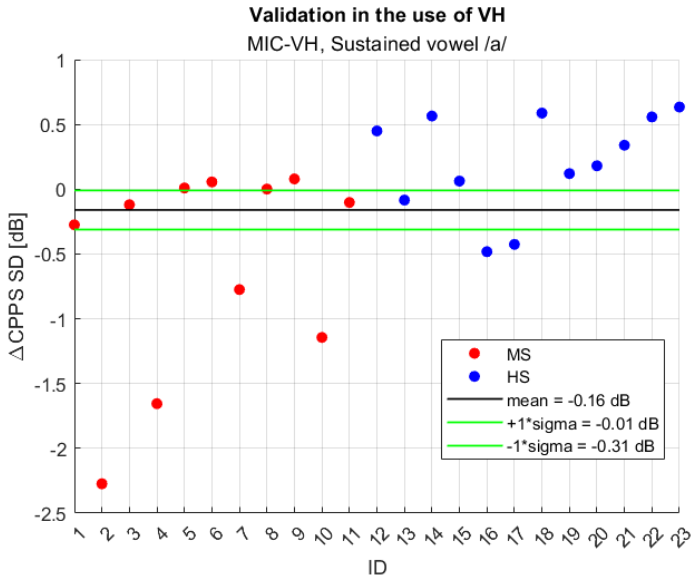

Figure 2.8: Example of delta values between the parameters extracted from in-air and VH for each subject in the case of sustained vowel /a/ task

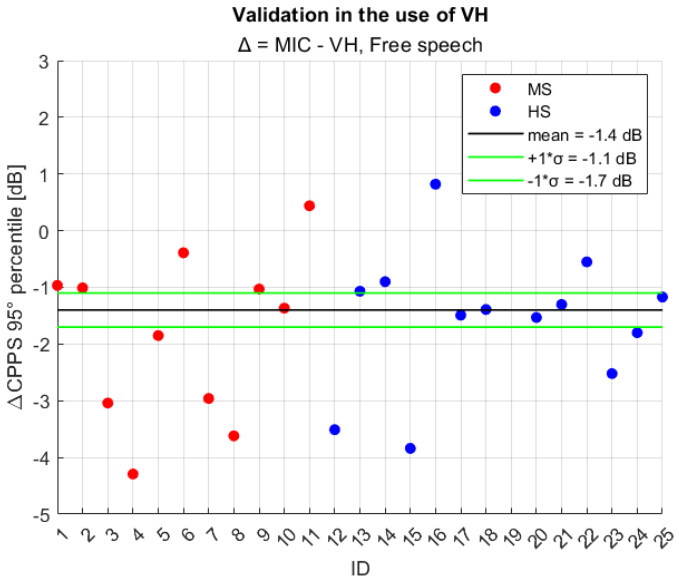

Figure 2.9: Example of delta values between the parameters extracted from in-air and VH for each subject in the case of free speech task

2.2.2 Other parameters acquired with VH device

In this part of the thesis the definitions of the vocal parameter Sound Pressure Level (SPL), and the environmental one, Background Noise Level (BNL), are provided. These two parameters are rendered available by VH device, but are not used for its validation (presented in section 2.2.1). Concerning the sustained vowel /a/ task, the SPL and BNL parameter are not returned by the device; regarding the balanced and free speech task, these two measures are downloadable from VH, however the validation of this device can be performed with the data in common with both microphones, and only f_0 and CPPS parameters (returned by both) can be compared. Furtheremore, a calibration of the in-air microphone (used to record the subjects in this study) can be performed, in order to create a function that converts the RMS values of the .wav files (from the in-air microphone) into the amplitude in dB corresponding to the SPL parameter resulting from the VH device. By means of this characterization for the in-air microphone, it follows that in the case of balanced and free speech task, the comparison between the RMS value (extracted from the microphone in air) and the SPL parameter (returned by VH) can be achieved. The difference between these parameters is not introduced in this work, since the calibration of the microphone in-air is not conducted. To better enable the visualization of the different parameters treated in this thesis, a reference is made to the diagram introduced in figure 2.1.

Background Noise Level BNL

Changes in voice production can be induced by environmental factors, such as noise level. The Background Noise Level (BNL), also called residual noise, is defined as an unwanted sound, emitted from outside the building where the recordings take place and those generated directly inside it. Most environmental sounds are made up of a complex mix of many different frequencies. The audible frequency range is normally considered to be $(20 \div 20000)$ Hz for young listeners with unpaired hearing. However, human hearing systems are not equally sensitive to all sound frequencies and, to compensate this, various types of filters or frequency weighting are used to determine the relative strengths of frequency components making up a particular environmental noise. The A-weighting is most commonly used and weights lower frequencies as less important than mid- and higher- frequencies [19]. The VH device returns four descriptive statistics for the background noise level parameter, all expressed in dBA, which are: BNL_{LAF50}, BNL_{LAF75}, BNL_{LAF90}, BNL_{Leq}. In this analysis the background noise activity level is measured at the same time as the subjects' voice recording and, is evaluated as the A-weighted level that exceeds for 90% the considered time (BNL_{LAF90} in dB). In addition, intelligibility is the percentage of words correctly understood by the listener, compared to the total number of sentences emitted by the speaker. Noise and poor acoustic characteristics

contribute to reduce intelligibility; the presence of an high background noise level in an environment, where communication is the key, often triggers the *Lombard effect*, which is the involuntary tendency of speakers to increase their voice level as the noise level increases in order to improve intelligibility of the speech signal while speaking in loud noise. The typical slope for the Lombard Effect is expected in the range $(0.3 \div 0.6)$ dB in voice increase for each dB of increase in the mean value of the A-weighted noise level distribution above 50 dBA [20].

Sound Pressure Level SPL

The VH device evaluates the speech Sound Pressure Level SPL parameter, which represents the local deviation from the ambient air pressure caused by a sound wave. The SPL is a logarithmic measure of the effective pressure of sound relative to a reference value; the commonly used reference sound pressure in air is 20 μPa, which is considered the threshold of human hearing. ISO 1999 defines sound pressure level (L_p) by the following formula:

$$L_p = 10 \log(\frac{p}{p_0})^2 \tag{2.23}$$

where, p is the sound pressure in Pascal and reference sound pressure p_0 is 20 μPa, in accordance with ISO 1683. The same Eq (2.23) can be used to determine the A-weighted sound pressure level L_{pA} where instead of p, the A-weighted sound pressure p_A in Pascal is used. Because of large sound pressure amplitude changes, the sound pressure level in decibels (L_p) is used rather than Pascal units. In the decibel scale, audible sounds range from 0 dB, the threshold of hearing, to over 130 dB, which is the threshold of pain. With more detail, a range of 0 to 40 dB is considered quiet to very quiet, while 60 to 80 dB is generally described as noisy. The SPL parameter returned by VH device is initially referred to a distance d_0 of 22 cm, which is the distance with respect to the subject's mouth set by the spacer inside the DAP unit of VH. With the objective of obtaining a value more comparable to the data in literature, the SPL value is expressed at a distance d equal to 1 m. The following expression is used to transform the SPL value:

$$\begin{aligned} SPL_d &= SPL_{d_0} + 20 \log_{10}(\frac{d_0}{d}) = \\ &= SPL_{d_0} + 20 \log_{10}(\frac{0.22}{1}) = SPL_{d_0} - 13{,}15 \end{aligned} \tag{2.24}$$

The SPL value at a distance of 1 m is calculated with the SPL value at a distance of 22 cm and by operating a subtraction with a constant value equal 13,15 dB. Although, VH device returns five descriptive statistics for the sound pressure level parameter all measured in dB, which are SPL_{mean}, SPL_{median}, $SPL_{5,\text{prc}}$ and $SPL_{95,\text{prc}}$ and SPL_{std}, in this analysis only SPL_{mean}, SPL_{median} and SPL_{std} are considered.

2.2.3 Intra-class and inter-class evaluation of VH parameters in long-term assessment of fatigue

The diffusion of long-term monitoring, instead of in-clinic short-term measurements, has provided significant parameters that, differently from average measures, are able to detect patients with abnormal vocal behavior related to voice disorders [12]. In this part of the study, other methods that lead to the discrimination between pathological and healthy subjects are described. Effort is focused on assessing the fatigue experienced by subjects during long-term acquisition rather than short-term (i.e., balanced and free speech task which last approximately one minute), the last considered as a sort of "baseline" (i.e., first instants of the monitoring). This comparison between the two classes is important especially when fatigue is often considered the most debilitating symptom for patients with MS, leading to loss of employment and impairment of activities of daily living [21]. As highlighted in Chapter 1 (section 1.1), in MS patients vocal fatigue and vocal breaks are more common than hoarseness [2]. To achieve this goal, only data acquired with VH are used in the current analysis; in particular, the data from the short term-assessment and the data from long-term monitoring are downloaded. The length of the long-term assessment is not fixed, but varies between the subjects; a minimum duration of 95 minutes is observed among all subjects (both MS and HS). The short-term data are the same already used during the validation of VH (presented in section 2.2.1) for the balanced/free speech task, but in this case also BNL_{LAF90} and SPL are added to fulfil this comparison (as reported in figure 2.10). Regarding the parameters estimated by VH device for long-term assessment, two types of files (which are downloadable through a proper script in Matlab (R2022b) environment) are rendered available. In one case, the parameters summarized in figure 2.11 are updated with a time interval of approximately 75 s (this file will be referred to as *tab file*). In the other type of file, the estimated parameters are only two (in figure 2.12) and are updated with a time interval of approximately 46 ms (that is referred to as *46ms file*). For this experiment, the *tab file* is used, since (as indicated in figure 2.11) it provides the same parameters in common with the short-term evaluation (figure 2.10).

BALANCED/FREE SPEECH TASK
Fundamental frequency, f_0 (Hz)
Smoothed Cepstral Peak Prominence, CPPS (dB)
Backgound Noise Level (90°perc), BNL LAF90 (dBA)
Sound Pressure Level, SPL (dB)

Figure 2.10: Parameters extracted from VH device for the short-term evaluation, used to perform the comparison with the long-term assessment

LONG-TERM EVALUATION (tab file)
Fundamental frequency, f_0 (Hz)
Smoothed Cepstral Peak Prominence, CPPS (dB)
Backgound Noise Level (90°perc), BNL LAF90 (dBA)
Sound Pressure Level, SPL (dB)

Figure 2.11: Parameters extracted from VH device (every 75 s) for the long-term evaluation, used to perform the comparison with the short-term assessment

LONG-TERM EVALUATION (46ms file)
Fundamental frequency, f_0 (Hz)
Sound Pressure Level, SPL (dB)

Figure 2.12: Parameters extracted from VH device (every 46 ms) for the long-term evaluation, used to perform the comparison with the short-term assessment

After downloading the data for both types of assessment, 75 s intervals that are characterized by a Voicing Time Percentage (%) value lower than 5% are removed, since the value of the parameters in this condition is considered not reliable. The idea is to compute the mean of the parameters (which is performed per column, since each column represent a parameter) for each subject under analysis. Four matrices are obtained (two for the long-term assessment relative to HS and MS, and two for the short-term evaluation relative to HS and MS) consisting of one row (i.e., mean value) for each subject under analysis. A proposal to assess fatigue is conducted with the use of differences, indicated with the greek small letter δ, between the parameters extracted from the long-term and the short-term monitoring. As a result, delta values for each of the metrics (columns) are obtained. To better clarify this concept, an example is shown in figure 2.13 that derives from the application of the Eq. (2.25) for all the parameters in common with the two assessments:

$$\delta = Long\,term\,recording - Short\,term\,recording \tag{2.25}$$

Having obtained delta values for each parameter, also their mean and standard deviation is considered. Red elements are related with patients with MS and blue ones are correspondent to healthy subjects. In addition, to better detect the dispersion of the values of delta, the standard errors of the mean are obtained, as a ratio between the standard deviation and the root square of the total number of subjects in each class and ,at the end, the value of the confidence interval is set to $\pm 1 \cdot \sigma$. In this sense, the two classes of subjects (HS and MS), their mean value of delta (reported as a black line for both classes) and confidence interval (indicated as the two red lines and the two blue lines for MS and HS respectively)

are kept separated. If the red band is lower than the blue band, it demonstrates that patients have a delta value (i.e., the difference between long-term and short-term monitoring) of the parameter considered, that is lower than that of healthy subjects. Moreover, it is important to assert that, if the two bands do not appear well separated, then, there is no significant difference in the behavior of the two classes with regard to the fatigue experienced. To better differentiate these bands, a suggestion proposed is to increase the data-set so as to narrow and make these intervals more distinguishable.

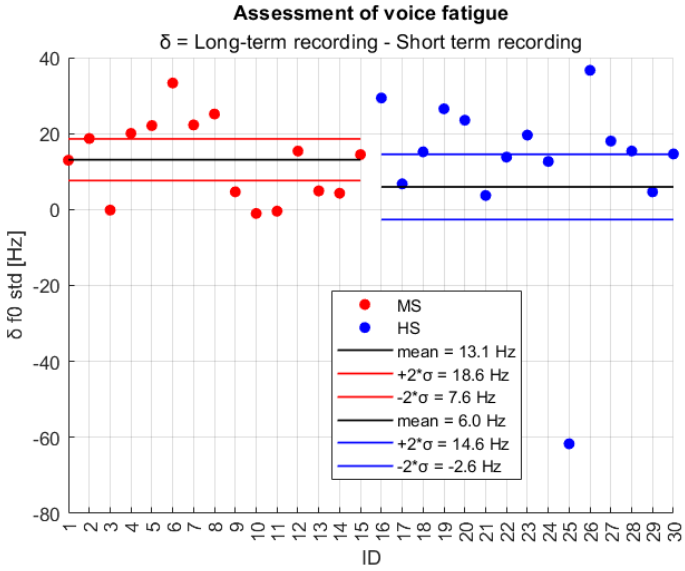

Figure 2.13: Example of delta values between the parameters extracted from the long-term and the short-term assessment for each subject in the case of balanced and free speech task

In the second part of this investigation, a qualitative analysis of the long-term monitoring is carried out using the *tab files* for all the involved subjects. These files, processed by VH device, are downloaded using a specific script in Matlab (R2022b) environment, and the parameters of interest are selected. Then, a selection of voiced and unvoiced frames is computed, and when the PPT results in the considered frame is lower than 5%, the parameters are set to zero, with the only exception of BNL_{LAF90} parameter; it is important to preserve the value of BNL_{LAF90}, regardless of whether the subject is talking or not, since the noise level is always present in any environment. In addition, being the SPL parameter initially referred to a distance from the subject's mouth d_0 of 22 cm (as reported in section 2.2.2), in order to operate the conversion at the distance of 1 m (which is most widely used in literature), the Eq (2.24) is used.

The extracted parameters are represented over time (as shown in the example in figure 2.14). In the existing literature, several works deal with in-field long-term monitoring of voice, but there is a lack of longitudinal studies that assess voice parameters modifications accounting for environment background noise level [12]. Since a strong correlation of data with the noise level in the environment is expected, these parameters are graphically presented with reference to BNL_{LAF90}. Having obtained these representations, a removal of unvoiced frames (which have been previously set to zero value) is carried out for all the parameters, in order to create different maps. An example is shown in figure 2.15 where the correlation between SPL_{mean} and BNL_{LAF90} parameter is depicted. In this sense, the idea is that, if the environment is characterized by a loud background noise level, then, one strategy that subjects usually perform in order to prevail, is to raise their speech intensity of voice (i.e., the SPL parameter). Three different correlations are represented at this point, that are SPL_{mean}-BNL_{LAF90}, $f_{0,\text{mean}}$-BNL_{LAF90} and SPL_{mean}-$f_{0,\text{mean}}$. Red elements are always related with MS patients and blue ones are associated to HS group. Additionally, using the Matlab command *polyfit* (which returns the coefficients for a polynomial of degree n that is a best fit for the data) and *polyval* (which evaluates the polynomial), a linear model (i.e., the degree of the polynomial is 1) that fits the data is plotted (as reported in figure 2.15). Referring to the example for SPL_{mean}-BNL_{LAF90} correlation, the regression line allows to observe the subject's ability to raise the intensity of voice depending on the noise level. For each map, along with the regression model, the correlation coefficient R^2, which expresses the goodness of the model, is also determined. An inter-class (i.e., among the two classes) and an intra-class (i.e., within the class) analysis is done; these three different types of correlation maps are computed both for each subject (in order to observe a difference inside each group) and by class (to describe the difference between MS and HS classes). Regarding the comparison conducted within the class, for each subject the three parameters of interest (BNL_{LAF90}, SPL_{mean} and $f_{0,\text{mean}}$) are saved, and two matrices are

created, containing all the data of the two groups. Finally, the already mentioned maps (SPL_{mean}-BNL_{LAF90}, $f_{0,\text{mean}}$-BNL_{LAF90} and SPL_{mean}-$f_{0,\text{mean}}$) are plotted, reporting all the data available for each class (as shown in figure 2.16 in the case of the HS class for the $f_{0,\text{mean}}$-BNL_{LAF90} correlation). By acquiring the same map, it is possible to make a comparison between individual subjects inside their class or between the two classes, e.g., given that MS patients suffer from hypophonia, it is expected that, in the long-term acquisition, while an healthy person is able to increase the level of SPL as the BNL increases, in the case of a pathological subject, this either does not occur or occur less markedly. Also, the SPL_{mean}-$f_{0,\text{mean}}$ map present a specific trend for the healthy class: if the speech intensity increases, then, the fundamental frequency also usually increases (for HS group) and the map is characterized by an asymmetrical shape with a tip toward the upper right corner. Before obtaining these maps, it is important to report that, the search and the subsequent elimination of outliers is performed; with more detail, subjects with a BNL_{LAF90} value equal 0 dBA (i.e., the acquisition does not occur correctly) and a fundamental frequency value higher than 300 Hz or lower than 60 Hz (i.e., not significant), are not considered in this analysis.

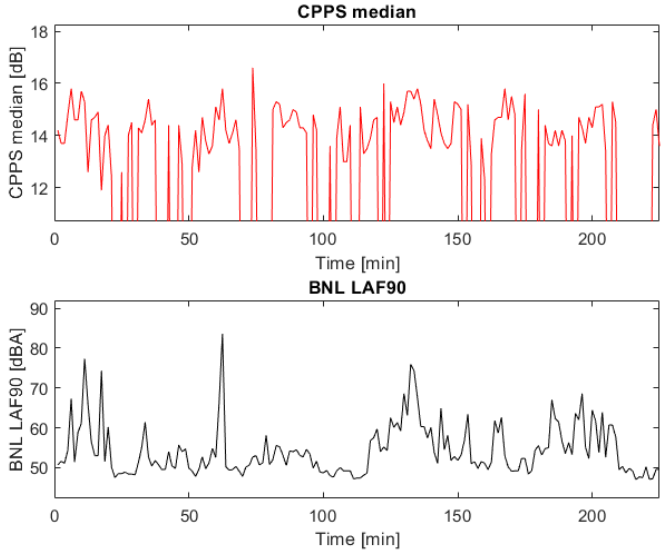

Figure 2.14: Example of $\text{CPPS}_{\text{median}}$ parameter over time with reference to the BNL_{LAF90} value present in the environment for one patient

In addition, considering the SPL_{mean}-BNL_{LAF90} correlation (shown in figure 2.15), a model can be derived, with the idea of compensating the speech intensity level for the effect of noise in the environment. Changes in voice production can be induced by environmental factors, such as the noise level; considering that, subjects

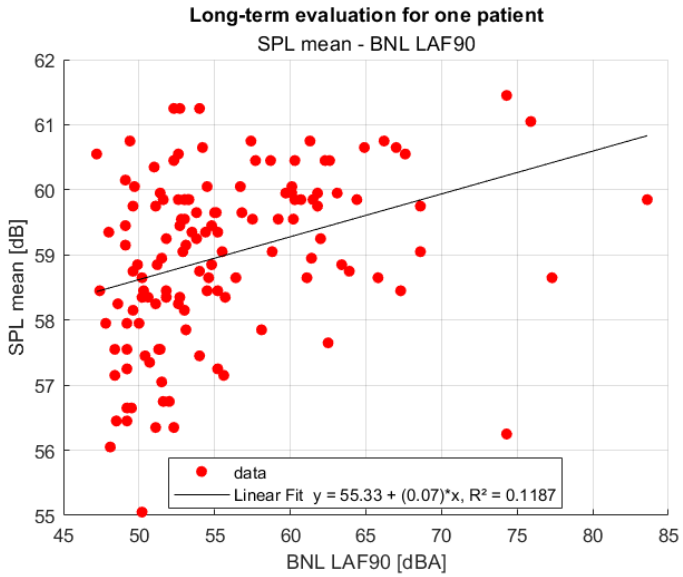

Figure 2.15: Example of correlation between the SPL_{mean} and the BNL_{LAF90} parameter for one patient

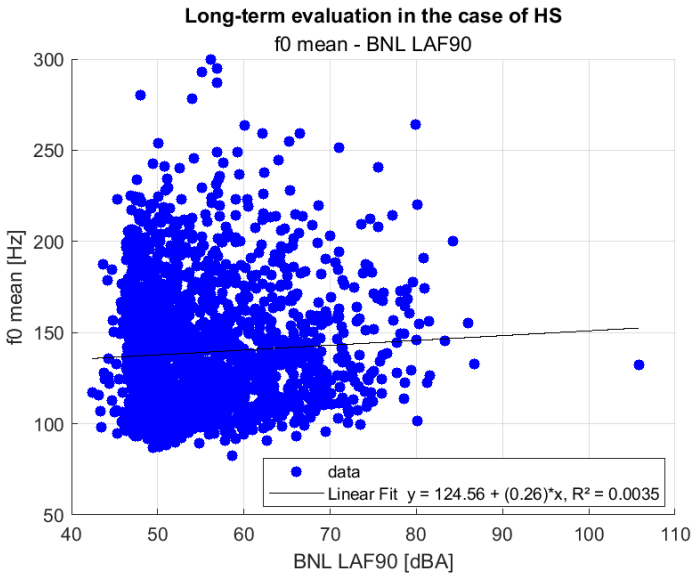

Figure 2.16: Example of correlation between the $f_{0,\text{mean}}$ and the BNL_{LAF90} parameter for HS class

raise the speech intensity level of voice, as a function of the noise present in the environment, a correction of the SPL_{mean} value with respect to the BNL_{LAF90} can

be performed. This compensation is done by performing a difference between the SPL_{mean} parameter and the angular coefficient of the regression line (that models the SPL_{mean}-BNL_{LAF90} correlation), multiplied by the BNL_{LAF90} parameter. The example (in figure 2.15) shows that the SPL_{mean} value increases by 0.07 dB each dBA of noise (i.e., 0.07 dB/dBA). As a result of the correction, the SPL_{mean} parameter over time is plotted (in figure 2.17), before (indicated in red) and after (in magenta) the BNL_{LAF90} correction. The patient shows an initial value of SPL_{mean} over time (i.e. the intercept at time zero) before BNL_{LAF90} correction, equal to about 58.0 dB and increases the speech intensity by 0.01 dB/min. The initial value of SPL_{mean} over time is corrected by subtracting the effect of noise and, as result, in an hypothetical condition without noise, the subject vocalizes at an initial SPL_{mean} value of about 54.4 dB and the slope (i.e. the angular coefficient of the regression line modelling SPL_{mean} over time) does not change over time, demonstrating that the subject does not show fatigue.

Eventually, to better observe differences in the slope values of the SPL_{mean} corrected between all the subjects, another analysis is performed. For each subject (HS and MS), the values of slope of the regression line (that models the SPL_{mean} corrected over time), are extracted and represented (figure 2.18). In addition, for each class the mean value (associated with the black horizontal line) and the standard deviation of the slope values are calculated. To better understand the dispersion of the slope values, the standard errors of the mean are obtained for both classes and the resulting value of the confidence interval is set to $\pm 1 \cdot \sigma$. The confidence intervals (i.e., the two red lines and the two blue lines) correspond respectively to MS patients and HS. By observing a separation between these two bands, a significant difference with respect to the slope values between the groups is found.

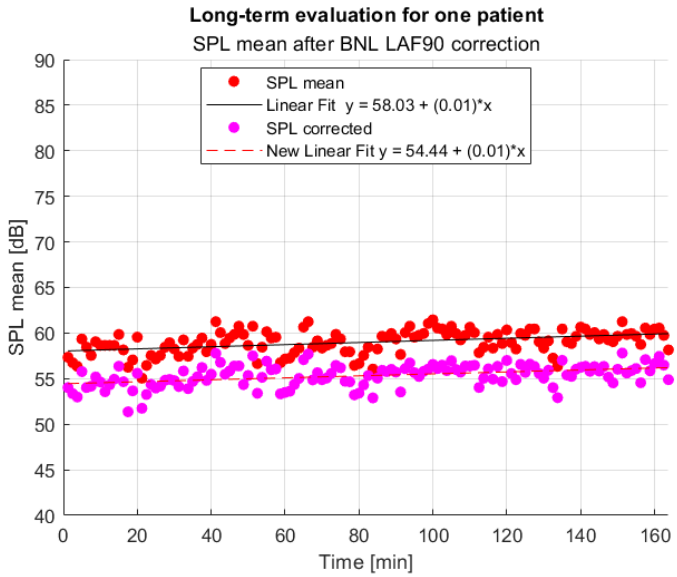

Figure 2.17: Example of SPL_{mean} compensation with respect to the BNL_{LAF90} over time for one patient

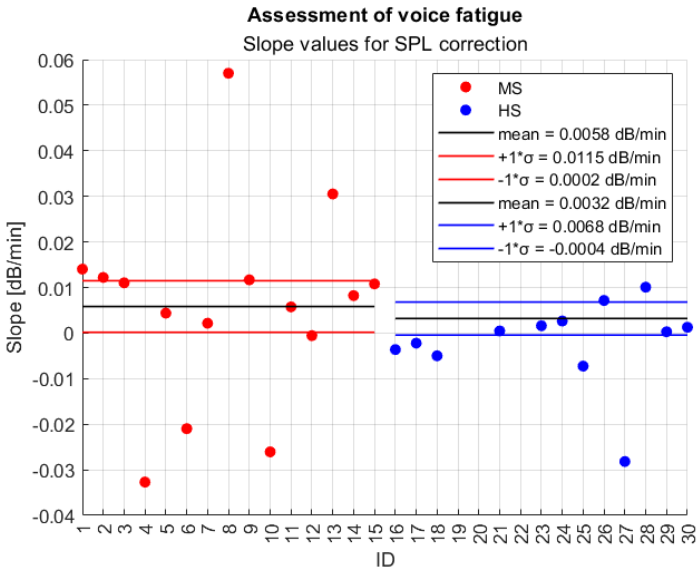

Figure 2.18: Representation of the extracted slope values (dB/min) of the regression line modelling the SPL_{mean} parameter compensated with respect to the BNL_{LAF90} over time for each subject

In addition, to assess the variability of the SPL_{mean} parameter over time (not

considering BNL_{LAF90} correction in this case) and to notice differences between the classes, the SPL_{std} parameter in dB (measured by VH device) for each subject is extracted and the mean value of this measure for the long-term monitoring, is calculated. For each class the mean and the standard deviation of the SPL_{std} parameter is obtained (these types of representations have already been presented for different parameters); also, in this case, the mean value of the SPL_{std} parameter is indicated (for both classes) with an horizontal black line, while the confidence interval (which is set to $\pm 1 \cdot \sigma$) is associated with the two red lines and the two blue lines for MS and HS respectively. The example (shown in figure 2.19) expresses the speech intensity variability for each subject. Similarly, the same steps can be performed on the $f_{0,\text{std}}$ parameter in order to evaluate the fundamental frequency variability between the classes.

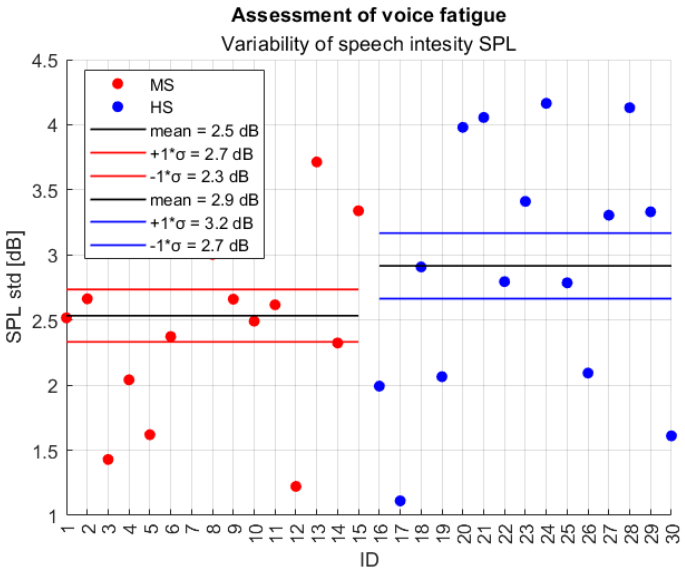

Figure 2.19: Example of the mean value of SPL_{std} parameter for each subject for long-term monitoring

2.2.4 VH as aid to quantify vocal exposure

In order to investigate the effects of prolonged or excessive voice use, it is important to properly quantify the amount of vocalization and three factors can be considered determinants that are the duration of voicing, vocal intensity and fundamental frequency. Vocal load is defined as a combination of prolonged voice use and additional factors, such as elevated phonation frequency and high sound pressure level [12]; while, vocal effort is a physiological magnitude that accounts for changes

in voice production induced by the distance from the listeners, noise and the physical environment [22]. In recent years, effort is focused in measuring the amount of voicing performed by speakers over time and devices, designated as voice accumulators, are used to determine parameters, such as speech intensity (measured as sound pressure level, SPL), fundamental frequency f_0 and phonation time. Prolonged vocal use can be considered as a problem of exposure of vocal folds to vibration, and although this vibration is self-induced, it resembles exposure to sun rays or chemicals. The term of "vocal dose" has been introduced by Titze, Švec and Popolo [23] and is adopted for measures quantifying the amount of voicing. In this section, the definitions of five vocal doses and the factors that can potentially contribute to abnormal vocal behavior are provided.

Time dose D_t

The simplest vocal dose is the time dose expressed in seconds, often called "voicing time" or "vocal accumulation time", which quantifies the total time during which the vocal folds vibrate and it is defined as:

$$D_t = \int_0^{t_m} k_v \, dt \tag{2.26}$$

where t_m is the total measurement time and k_v is the voicing unit step function (which is 1 when the frame is voiced and 0 if unvoiced). It is possible to relate the time dose to the Voicing Time Percentage (PPT), which can be calculated operating a ratio between the time dose and the total measurement time and then, multiplying it by 100.

Cycle dose D_c

It quantifies the total number of oscillatory periods in cycles, completed by the vocal folds over time and it can be expressed by Eq. (2.27):

$$D_c = \int_0^{t_m} k_v f_0 \, dt \tag{2.27}$$

where f_0 is the fundamental frequency of the vocal folds oscillation (Hz).

Distance dose D_d

In order to account also for amplitude vibration, the distance dose is introduced, which measures the total distance accumulated by the vocal folds in a cyclic path during vibration. Its definition is expressed in meters as indicated in:

$$D_d = 4 \int_0^{t_m} k_v A f_0 \, dt \tag{2.28}$$

where A is the amplitude of the vocal folds. The number four in Eq. (2.28) is explained because the vocal folds theoretically travel at a distance of four times the amplitude within a cycle. Since the amplitude of the vocal folds changes with the vocal intensity, D_d accounts for both the intensity and the fundamental frequency in voicing. In addition, the amplitude of the vocal folds is very difficult to measure and in order to overcome this problem, an approximation using existing normative data is performed. The amplitude A can be calculated using the empirical rules in Eq. (2.29):

$$A = 0.05L_0[(P_L - Pth)/P_{th}]^{\frac{1}{2}} \tag{2.29}$$

where L_0 is a reference vocal folds length (which is 0.016 m for males and 0.01 m for females), P_L is the lung pressure and P_{th} is the phonation threshold pressure. The rule for the definition of P_{th} is presented in Eq. (2.30):

$$P_{th} = 0.14 + 0.06(f_0/f_{0N})^2 \tag{2.30}$$

where f_0 is the fundamental frequency and f_{0N} is a nominal fundamental frequency (which is 120 Hz for males and 190 Hz for females). The empirical expression for the lung pressure P_L needed to determine the distance dose is calculated as:

$$P_L = P_{th} + 10^{(SPL-78.5)/27.3} \tag{2.31}$$

The Eq. (2.31) is derived for the SPL parameter measured at the distance of 50 cm from the mouth.

Energy dissipation dose D_e

It takes into account the factor of thermal agitation of tissue inside the vocal folds and measures the amount of heat produced during vibration as calculated in Eq. (2.32):

$$D_e = \frac{1}{2}\int_0^{t_m} k_v \eta (A/T)^2 \omega^2 \, dt \tag{2.32}$$

The energy dissipation dose is measured in joules/m^3 and in Eq. (2.32) η is the shear viscosity of the vocal folds tissue evaluated in Pascal $\cdot$ s, T is the vertical thickness of the vocal folds expressed in meters and $\omega = 2\pi f_0$ is the angular frequency of the vocal folds vibration in rad/s. Both shear viscosity η and vertical thickness T are approximated from the frequency f_0 of voice using the empirical rules:

$$\eta = \begin{cases} 5.4/f_0 \; for \; males \\ 1.4/f_0 \; for \; females \end{cases} \tag{2.33}$$

$$T = \begin{cases} \frac{0.0158}{1+2.15(f_0/120)^{1/2}} \; for \; males \\ \frac{0.01063}{1+1.69(f_0/190)^{1/2}} \; for \; females \end{cases} \tag{2.34}$$

Radiated energy dose Dr

The fifth dose presented is not a measure of exposure to the vocal folds, but rather a potential sound exposure to a listener and it quantifies the total energy radiated from the mouth in joules over time as expressed in Eq. (2.35):

$$D_r = 4\pi R^2 \int_0^{t_m} k_v 10^{(SPL-120)/10}\, dt \tag{2.35}$$

where R is the distance from the mouth (that in this case corresponds to 0.5 m) at which the SPL of voice is recorded. Using these definitions and empirical rules, all the doses can be derived for a specified measurement time t_m, by extracting three basic parameters of speech, that are k_v (i.e., voicing/unvoicing parameter), f_0 and SPL. The time, cycle and radiated energy doses are the true doses for the person measured, whereas the distance dose and the dissipated energy dose are approximations based on typical data for male and female vocal folds amplitudes, thickness and viscosities [24].
To perform this investigation, the long-term acquisition (performed with VH device) indicated as *46ms file* is used. This file for each subject is downloaded through a specific script in Matlab (R2022b) environment and returns the values (two columns) respectively of f_0 and SPL parameter with a time interval of 46 ms. As already mentioned, the SPL parameter is referred to a distance from the subject's mouth of 22 cm. Considering the *46ms file*, the already seen Eq. (2.24) used to refer the SPL parameter to a distance of 1 m is considered. Then, having obtained the SPL value at a distance of 1 m with the use of Eq. (2.24) the SPL at 50 cm as indicated in [24] by Titze, Švec and Popolo is determined. The voicing unit step function is created searching for voiced or unvoiced frames (i.e., frames having an SPL or f_0 value equal to zero). The SPL, f_0 and k_v values of each 46 ms frame length, are used for calculating the vocal doses for each subject under analysis. Also, the equivalent SPL value at 1 m from the speaker's mouth (i.e. $SPL_{eq,1m}$) is estimated for each subject, which expresses the speaker's vocal effort according to ANSI S3.5-1997 standard [25]. $SPL_{eq,1m}$ is calculated as the average of the voiced energy over all the frames, including the unvoiced ones, as indicated by Švec *et al.* [26] as follows:

$$SPL_{eq,1m} = 10\log_{10}\left(\frac{1}{N}\sum_{n=1}^{N}[k_v 10^{SPL_{1m}/10}]\right) \tag{2.36}$$

where N is the total number of frames in the analyzed segment of speech and, with inclusion of the k_v factor the energy in the unvoiced frames is set to zero, while for voiced frames is set to one. The accumulation of the different doses is calculated (figure 2.20 for the energy dissipation dose case) and, it is possible to observe that the dose values increase during the voiced passages, whereas they stay constant during unvoiced segments, as expected.

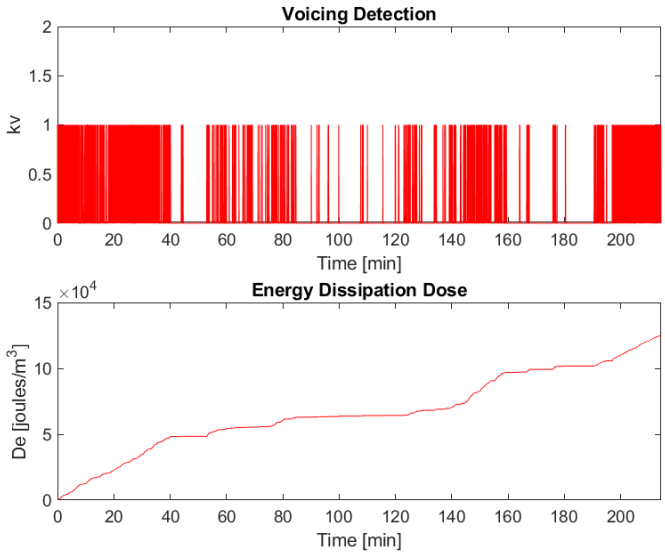

Figure 2.20: Accumulation of the energy dissipation dose and the correspondent voicing/unvoicing parameter over a 214-minutes segment of speech for a male patient

These results led to the difficulty that, unless during the acquisition, vocal effort is significant among the other subjects, then, there is the problem of being able to compare the five vocal doses. All subjects under analysis performed long-term monitoring ranging from 95 minutes up to 326 minutes. Therefore, in order to compare different subjects and compute an assessment at consistent times, a minimum available duration lasting 95 minutes common to all is considered. For each subject a matrix with a dimension of *1x5* is extracted containing the five vocal doses at the minute 95 (i.e., D_t at 95 min, D_c at 95 min, D_d at 95 min, D_e at 95 min and D_r at 95 min). Each matrix is then concatenated in order to perform a discrimination between the two groups of HS and MS. Then, the value of each dose at the minimum time chosen for each subject are represented (as in the example of the distance dose in figure 2.21); red elements are related to MS and blue ones to HS. Also, in this case, to better detect the dispersion of the vocal doses between the HS and MS subjects, the standard errors of the mean are obtained as a ratio between the standard deviation and the root square of the total number of subjects in each class and at the end, the value of the confidence interval is set to $\pm 1 \cdot \sigma$. In this sense, the two classes of subjects (HS and MS), their mean value of delta (reported as a black line for both classes) and confidence interval (indicated as the two red lines and the two blue lines for MS and HS respectively) are kept separated. If, in the example relative to the distance dose, the red band is lower than the blue

band, it demonstrates that patients present a distance dose mean value (i.e., the total distance travelled by the vocal folds in an oscillatory path) that is lower than that of healthy subjects. Moreover, it is important to assert that if the two bands do not appear well separated, then, there is no significant difference in the behavior of the two classes regarding fatigue.

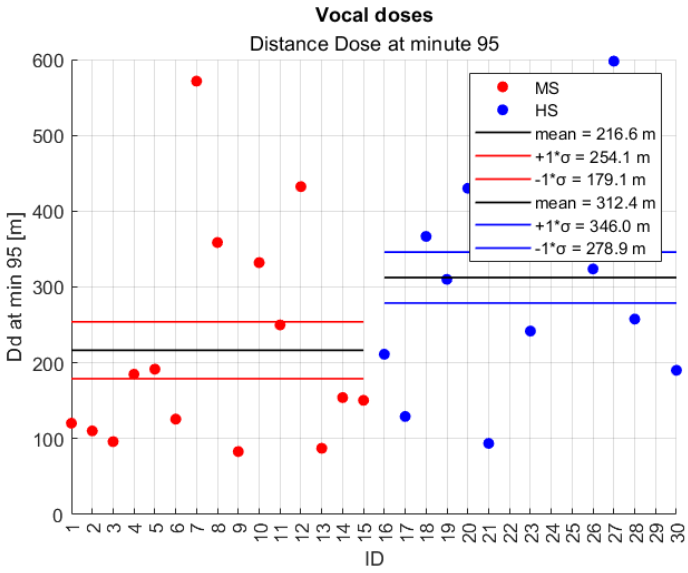

Figure 2.21: Example of distance dose values at the minimum time interval of 95 minutes for each subject

The time dose can be used for quantifying the duration of voicing, and the voicing percentages among various vocal activities or occupations. Furthermore, it can also be used as a normalization factor to obtain doses per second of vocalization. In this sense, another analysis performed is to weight the vocal doses with respect to the time dose. However, being all the vocal doses extracted for a time interval of 95 minutes, the result is the same if the doses are referred to the time dose, instead of the PPT (since the considered interval of time is the same). An example of the cycle dose represented in function of time dose is shown in figure 2.22: a strong correlation is expected between D_c and D_t, since the only variability is introduced by f_0 parameter and a lower correlation between D_d and D_t doses, the latter motivated by the fact that, in the definition of distance dose empirical rules are introduced. Moreover, in vocal doses where no significant dependence with the time dose is observed, then, data can be represented without operating a weighting with respect to the interval of phonation. Additionally, it should be noted that, a limiting factor in this analysis is having reduced the monitoring to a time interval of 95 minutes as the minimum available interval common to all subjects. This

choice, is also motivated by the fact that the data-set is not wide. The effect of fatigue experienced by subjects is better observable in long-term monitoring, and the objective is always discriminating between healthy and patients. With this purpose, the same investigation is performed by eliminating those subjects that presented a long-term evaluation which lasts less than the minimum time interval of duration and increasing the time interval to 156 and then to 200 minutes of duration, in the latter case removing from the study a total of four subjects.

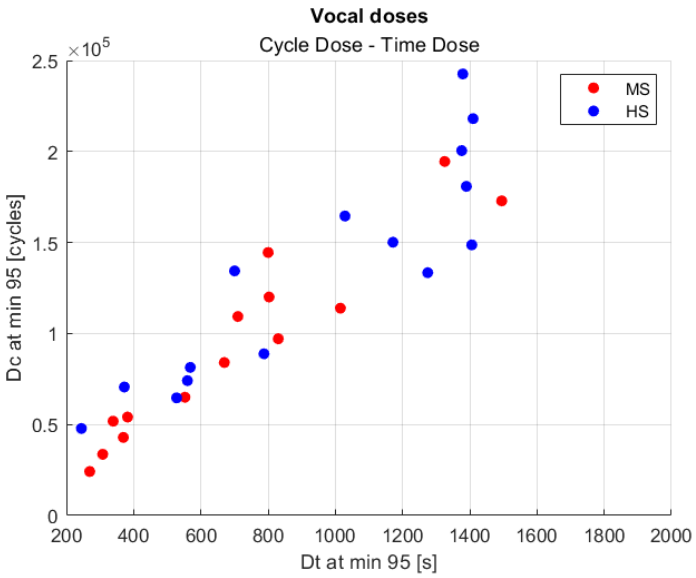

Figure 2.22: Example of cycle dose values weighted with respect to the time dose at the minimum time interval of 95 minutes

Chapter 3

Results

In this chapter, the results obtained are presented and discussed. The findings resulting from the in-air microphone proposed procedure are shown; the figures in the subsequent sections summarize the combination of features with the best accuracy obtained in the training phase of Feature Selection (FS) and the best accuracy values obtained after the 5-fold cross-validation of the Logistic Regression (LR) model. In addition, a procedure based on the confidence level of the probability returned by the LR model is proposed, in order to provide a more realistic evaluation of the classification performance. The last steps characterize the Vocal Holter (VH) device intended as a useful tool for assessing vocal health and validating its application in in-clinic consultation in comparison with the microphone in air. This comparison is carried out by calculating differences Δ between the parameters extracted from the two microphones. In addition, a proposal to assess fatigue is conducted with the use of differences δ between the parameters extracted from the long-term and the correspondent short-term monitoring. The parameters acquired with VH for the long-term monitoring are visualized over time and an analysis considering the effect of the background noise level in respect to voice production is executed. The final phases investigate five vocal doses as indicators of long-term vocal folds tissue exposure to vibration.

3.1 Logistic Regression results

A LR model is used to provide the most significant features that are good descriptors for the classification of subjects' voices as pathological or healthy. The LR algorithm provides a continuous probability p of belonging to the positive class; this probability is compared to a fixed threshold, which in binary classification, is equal to 0.5. The elements displaying a p higher than the threshold are assigned to the positive class, otherwise, are attributed to the negative class. The data-set provided is subdivided

between the negative class of healthy control group of subjects (HS, 16 subjects) and the positive class which is associated with the group of patients with Multiple Sclerosis (MS, 16 patients).

3.1.1 Feature-selection results

Feature Selection process is based on the evaluation of the performance of the LR model, which is trained using a different number of input features. The algorithm is developed in order to select a combination of k features (which ranges from 1 to 4) out of the number of features available for the task under consideration; as previously exposed in section 2.1.3, for the balanced and free speech task 47 features are evaluated, while in the case of sustained vowel /a/ there are 56 available features. About the combination of two, three or four features, only the ones that exhibited a correlation lower than 0.5 and a p-value of the evaluated correlation lower than 0.05, are accepted as input of the LR model. The algorithm, during the training phase, automatically indicates multiple combinations of features with the best classification performances which are used to validate the LR model. This research is conducted checking carefully for the combination of features with the highest accuracy value; in the cases where multiple feature sets provided the same maximum accuracy value, the validation is conducted for all sets, by selecting the combination with the highest AUC values among the ones with the highest accuracy. The figures in this section are relative to the training phase of the LR model and the results for all three tasks are presented. In order to better differentiate, the tables reporting the instruction "no validation" are related to the training phase of the algorithm; if in the tables indicate "5-fold cross validation", these are associated with the validation phase (covered in section 3.1.2). In the upper left corner of the tables the two classes involved in this work are presented; the negative class is associated with the healthy subjects and the positive class is related to the MS patients. Also, all tables provide the information about the analyzed task (sustained vowel /a/, balanced speech task and free speech task) and the "Features" consist in the outcomes of feature selection. As performance metrics, Area Under the Curve (AUC), Precision, Sensitivity, Specificity and Accuracy are exposed (their definition is reported in Appendix A.3).

Sustained vowel /a/

In figure 3.1 the combinations of features providing the best classification performance in the case of sustained vowel /a/ are listed. It is possible to notice that, in this task, the algorithm is not capable of performing the combination of four features together. In this case the highest accuracy value and AUC value in the training phase of the model is obtained with a combination of three features; in

particular as far as concerns the repetition of sustained vowel task the accuracy values do not exceeds the value of 85,2% by selecting the features Coefficient of Amplitude Variation (vAm), Harmonic to Noise Ratio (standard deviation) and Cepstral Peak Prominence Smoothed (skewness). It proves to be interesting the fact that in the case of vowel /a/ task, it's more evident the presence of perturbation parameters (described in section 2.1.2), such as vAm and Relative Average Perturbation. Generally looking at the selected parameters, it should be noted the recurrence of the parameter $\text{CPPS}_{\text{skewness}}$ for all the possible number of combinations. Also, it is important to highlight the fact that in all the reported cases, the results are lowered during 5-fold cross validation of LR model (as shown in section 3.1.2).

HS (0) vs MS (1)	SUSTAINED VOWEL /a/ NO VALIDATION				
Features	AUC	Precision	Sensitivity	Specificity	Accuracy
CPPS (skewness)	80,0%	88,9%	66,7%	93,3%	81,5%
RAP, CPPS (skewness)	84,2%	90,0%	75,0%	93,3%	85,2%
vAm, HNR (std), CPPS (skewness)	85,0%	83,3%	83,3%	86,7%	85,2%

Figure 3.1: Classification performance obtained without validation in sustained vowel /a/ task

Balanced speech task

For balanced speech task the metrics for classification performance are shown in figure 3.2. Looking at the selected parameters for this task, the most common ones are the harmonic frame ratio (V/uV), which express the percentage of harmoniousness over the voiced frames and the number of harmonic frames. In addition, CPPS (mode, 5° percentile, kurtosis) and HNR (mean, mode, 5° percentile) parameters are selected many times in the different combinations of features. Also, the first more evident difference in respect to the previous task of sustained vowel /a/ is the high number of selected combinations of features, in particular for combinations of three features. Moreover, it should be noted that, regarding the balanced speech case, the accuracy values exceed 90% and the best situation is trained with $\text{CPPS}_{\text{mode}}$ and V/uV. Another unusual characteristic highlighted in this task if compared to the others, is the presence of *gender* (i.e., "male" or "female") as selected parameter. In fact, gender in this work is considered a significant feature along with the other extracted parameters for classification and, together with $\text{CPPS}_{5,\text{prc}}$ and V/uV parameters, shows to be the combination offering the best performance also during validation phase of the LR model (as reported in the section 3.1.2). This result, drive the analysis toward the employment of fundamental frequency f_0 (instead of *gender*) in the above-mentioned feature combination; moreover, during the

validation phase, a test is conducted by selecting the parameters $f_{0,\text{mean}}$, $\text{CPPS}_{5,\text{prc}}$ and V/uV and the classification metrics of this combination is evaluated. The choice in using $f_{0,\text{mean}}$ is derived by considering that fundamental frequency (which is perceptually highly correlated with pitch), notably differs in men and women mainly due to anatomical variations [6].

HS (0) vs MS (1)	BALANCED SPEECH TASK				
	NO VALIDATION				
Features	AUC	Precision	Sensitivity	Specificity	Accuracy
HNR (mode)	83,0%	88,9%	72,7%	93,3%	84,6%
V/uV	84,2%	81,8%	81,8%	86,7%	84,6%
CPPS (mode), V/uV	100,0%	100,0%	100,0%	100,0%	100,0%
GENDER, CPPS (5° perc), V/uV	92,1%	90,9%	90,9%	93,3%	92,3%
f_0 (5° perc), CPPS (5° perc), V/uV	92,1%	90,9%	90,9%	93,3%	92,3%
HNR (mean), CPPS (5° perc), number harmonic frames	92,1%	90,9%	90,9%	93,3%	92,3%
HNR (mean), CPPS (kurtosis), number harmonic frames	92,1%	90,9%	90,9%	93,3%	92,3%
HNR (5° perc), CPPS (5° perc), CPPS (kurtosis), number harmonic frames	96,7%	91,7%	100,0%	93,3%	96,2%

Figure 3.2: Classification performance obtained without validation in balanced speech task

Free speech task

Concerning the results for free speech task, it is immediately noticeable the similarities with the just mentioned balanced speech task; although, the descriptive statistics in some cases are different, a recurrence of the same parameters selected in the reading case is evident, these are V/uV, CPPS (mode, 5° percentile, skewness and kurtosis) and HNR (standard deviation, range and 5° percentile). The number of combinations is slightly lower if compared to the balanced speech case. As far as concerns free speech task, the highest accuracy value is reached by selecting the features $\text{CPPS}_{\text{mode}}$ and V/uV, but there are also other cases where the accuracy exceeds 90%.

HS (0) vs MS (1)	FREE SPEECH TASK				
	NO VALIDATION				
Features	AUC	Precision	Sensitivity	Specificity	Accuracy
V/uV	81,8%	83,3%	76,9%	86,7%	82,1%
CPPS (mode), V/uV	100,0%	100,0%	100,0%	100,0%	100,0%
HNR (std), CPPS (5° perc), V/uV	96,2%	100,0%	92,3%	100,0%	96,4%
f_0 (5° perc), CPPS (skewness), V/uV	96,2%	100,0%	92,3%	100,0%	96,4%
HNR (range), HNR (5° perc), CPPS (5° perc), CPPS (kurtosis)	79,0%	73,3%	84,6%	73,3%	78,6%

Figure 3.3: Classification performance obtained without validation in free speech task

By looking at the AUC and accuracy values exposed in this section, it is possible to ascertain that the LR model performs satisfactorily during training phase and moreover, best results are obtained for balanced and free speech task in comparison with sustained vowel /a/ case, in both tasks by selecting the features $\text{CPPS}_{\text{mode}}$ and V/uV.

3.1.2 Best performance of validation phase

This section summarizes the best accuracy values obtained after the 5-fold cross validation of the LR model in the Classification Learner App in Matlab (R2022b). As described in section 2.1.4, the k-fold cross validation is aimed in avoiding over-fitting errors, by sub-dividing the data-set into subsets (five in our case) of equal size, and in using each subset at time in rotation to validate the model. The validation outcomes of the models presented in the previous section, related to the training phase of the LR model, are reported. Since the App returns for each model the confusion matrix relative to the validation phase, the performance metrics for all three tasks are obtained.

Sustained vowel /a/

The situations with the same highest accuracy value, up to 77,8% for the sustained vowel /a/ task are reported; these are obtained by selecting $\text{CPPS}_{\text{skewness}}$ parameter and feature combination comprising the parameters vAm, HNR_{std} and $\text{CPPS}_{\text{skewness}}$. In the following tables, the rows (i.e., the combination of features) with the highest classification performances are highlighted in yellow. Since there are multiple combinations of features with the same highest accuracy and AUC values, the model consisting of the lower number of selected features is considered; in this task, the highest classification metrics are obtained with by choosing $\text{CPPS}_{\text{skewness}}$ feature. In addition, in all three task the LR validated model of the combination of features showing the best performance (i.e. the combination highlighted in yellow) is saved with the use of "Generate Function" command (in the Classification Learner App), gaining the possibility to recreate the same model later in Matlab environment.

HS (0) vs MS (1)	SUSTAINED VOWEL /a/ 5-fold cross VALIDATION				
Features	**AUC**	**Precision**	**Sensitivity**	**Specificity**	**Accuracy**
CPPS (skewness)	76,7%	80,0%	66,7%	86,7%	77,8%
RAP, CPPS (skewness)	69,2%	70,0%	58,3%	80,0%	70,4%
vAm, HNR (std), CPPS (skewness)	76,7%	80,0%	66,7%	86,7%	77,8%

Figure 3.4: Classification performance obtained after computing 5-fold cross validation in sustained vowel /a/ task

Balanced speech task

The combination of features with the highest accuracy value is obtained by selecting *gender*, $\text{CPPS}_{5,\text{prc}}$ and V/uV, reaching up to 92.3%. This situation is considered the one returning the best classification metrics between all three task. The outcomes

of the investigation (anticipated in section 3.1.1), regarding the correlation between the parameters $f_{0,\text{mean}}$ and *gender* are described (figure 3.5); the $f_{0,\text{mean}}$ parameter (instead of *gender*) is considered together with $\text{CPPS}_{5,\text{prc}}$ and V/uV. The results obtained, are lowered in respect to the previous feature combination (*gender*, $\text{CPPS}_{5,\text{prc}}$ and V/uV), but still express how much this combination of features is significant in distinguishing between positive and negative classes.

HS (0) vs MS (1)	BALANCED SPEECH TASK				
	5-fold cross VALIDATION				
Features	AUC	Precision	Sensitivity	Specificity	Accuracy
HNR (mode)	83,0%	88,9%	72,7%	93,3%	84,6%
V/uV	84,2%	81,8%	81,8%	86,7%	84,6%
CPPS (mode), V/uV	85,5%	76,9%	90,9%	80,0%	84,6%
GENDER, CPPS (5° perc), V/uV	92,1%	90,9%	90,9%	93,3%	92,3%
f_0 (5° perc), CPPS (5° perc), V/uV	88,8%	83,3%	90,9%	86,7%	88,5%
HNR (mean), CPPS (5° perc), number harmonic frames	84,2%	81,8%	81,8%	86,7%	84,6%
HNR (mean), CPPS (kurtosis), number harmonic frames	79,7%	80,0%	72,7%	86,7%	80,8%
HNR (5° perc), CPPS (5° perc), CPPS (kurtosis), number harmonic frames	79,7%	80,0%	72,7%	86,7%	80,8%
f_0 (mean), CPPS (5° perc), V/uV	87,6%	90,0%	81,8%	93,3%	88,5%

Figure 3.5: Classification performance obtained after computing 5-fold cross validation in balanced speech task

Free speech task

It is noticeable (in figure 3.6) that, in free speech task the same best accuracy and AUC values are obtained in two cases with a combination of three features; these consist respectively in HNR_{std}, $\text{CPPS}_{5,\text{prc}}$, V/uV and $f_{0,5,\text{prc}}$, $\text{CPPS}_{5,\text{prc}}$, V/uV. Both features combinations reach the accuracy value of 89.3%, which is slightly lower than the previous in balanced speech case.

HS (0) vs MS (1)	FREE SPEECH TASK				
	5-fold cross VALIDATION				
Features	AUC	Precision	Sensitivity	Specificity	Accuracy
V/uV	77,9%	81,8%	69,2%	86,7%	78,6%
CPPS (mode), V/uV	82,3%	78,6%	84,6%	80,0%	82,1%
HNR (std), CPPS (5° perc), V/uV	89,5%	85,7%	92,3%	86,7%	89,3%
f_0 (5° perc), CPPS (skewness), V/uV	89,5%	85,7%	92,3%	86,7%	89,3%
HNR (range), HNR (5° perc), CPPS (5° perc), CPPS (kurtosis)	67,9%	64,3%	69,2%	66,7%	67,9%

Figure 3.6: Classification performance obtained after computing 5-fold cross validation in free speech task

It jumps to the eye the fact that, regarding the vocalizations performed in natural conditions (i.e., balanced and free speech task), both the models, achieving best performances in terms of discriminating healthy from pathological voices, are characterized by selecting the features $\text{CPPS}_{5,\text{prc}}$ and V/uV; the parameter that changes in the two task, in the case of the balanced speech task is *gender*, whereas, in the free speech case is HNR_{std} (as reported in figures 3.5 and figure 3.6).

The accuracy values for sustained vowel task are more than ten percentage points lower than those of balanced and free speech task. These results are in accordance with those reported in literature [14] [27]. It has been argued that asking subjects to produce sustained vowel seems to be somehow artificial and for that reason clinicians prefer running speech monitoring (i.e., balanced/free speech task) when they evaluate voice quality perceptually.

3.2 Realistic classification performance based on uncertainty evaluation

In this part of the chapter the results obtained with the method (described in section 2.1.5) relative to the Expanded Uncertainty $U(p)$ of the probability p returned by the LR model are exposed. As already mentioned, for each task the LR validated model is recreated by the command "Generate Function" in the Export section from the Classification Learner App. The generated Matlab (R2022b) code returns the classification model validated in the App, which can be used to extract predictors and responses in order to compute new performance metrics; this function receives as input the data-set with features and classes and as output a structure containing various fields with information on the trained classifier.
The standard uncertainty $u(p)$ of the LR validated model is estimated through the uncertainty propagation formula, which is implemented on the probability p returned by Eq. (A.2) (see in the Appendix A.3). All the β coefficients of the Eq. (A.1) are associated by an uncertainty (SE) value and a covariance value, both returned by the above-mentioned function. The formula is presented in Eq. (3.1):

$$\begin{aligned} u(p) &= \sqrt{J_\beta \cdot COV_\beta \cdot J_\beta^T} \\ J_{i,j}(\beta) &= \frac{\partial p_i}{\partial \beta_i};\ j \in [1...N_F + 1];\ i \in [1...N_S] \end{aligned} \tag{3.1}$$

where N_F is the number of considered features, N_S is the number of samples in the data-set, $J_{i,j}(\beta)$ is the Jacobian matrix of the model coefficients and COV_β is the variance-covariance matrix of the coefficients. In this analysis, the decision to pursue the computation of the expanded uncertainty only on models achieving best accuracy values during validation phase, is taken; these models (reported in section 3.1.2) are highlighted in yellow.

Sustained vowel /a/

Concerning the three repetitions of vowel /a/ task, the LR validated model with the highest accuracy value is obtained by selecting the parameter $\text{CPPS}_{\text{skewness}}$. Since (in this case) the number of selected features k is equal to one, the uncertainty

formula in (3.1) is applied with the contribution of two sensitivity coefficients and two covariances. The probabilities, returned by the LR validated model, without considering yet the expanded uncertainty from the Classification Learner App are shown (in figure 3.7) using blue squares for the negative class of healthy patients, and red circles for the positive class of MS patients.

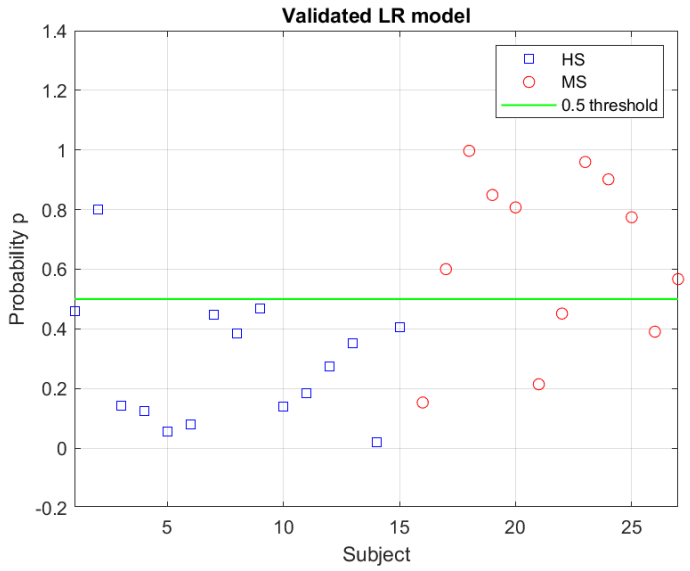

Figure 3.7: Probabilities returned by the LR validated model without the implementation of the expanded uncertainty for sustained vowel /a/ case

The expanded uncertainty $U(p)$ is obtained by multiplying the uncertainty $u(p)$ by a coverage factor of 2; the confidence interval for each probability value of the model is obtained and graphically represented with the use of errors bars (figure 3.8). The discrimination probability threshold, equal to 0.5, is indicated as a thick green line; since subjects with $p > 0.5$ are assigned to the MS class, the total number of wrong classification is six (two false positive and four false negative) as reported in the confusion matrix (in figure 3.9).

It is noted that, for six subjects of the negative class and four subjects of the positive class, the confidence interval includes the discrimination probability threshold, thus making questionable the classification of these subjects. Hence, the decision to tag these subjects as "non-classifiable" is taken, and the subsequent estimation of the classification performance metrics (by excluding them), is performed. In the considered case, the "non-classified" subject are eleven out of a total number of subject equal to twenty-seven. In order to better understand, the effect of the "non-classified" subjects on the overall performance of the classifier, new classification metrics such as Realistic Accuracy ($Accuracy_R$) and Fraction of

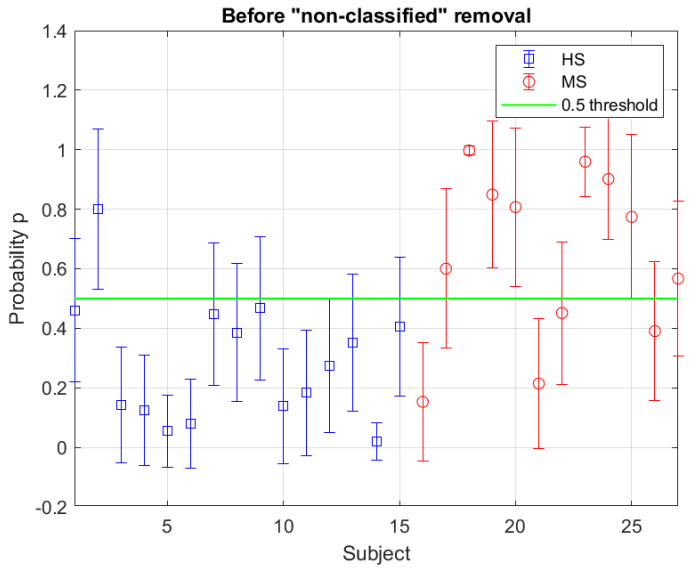

Figure 3.8: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for sustained vowel /a/ case before the removal of "non-classified"

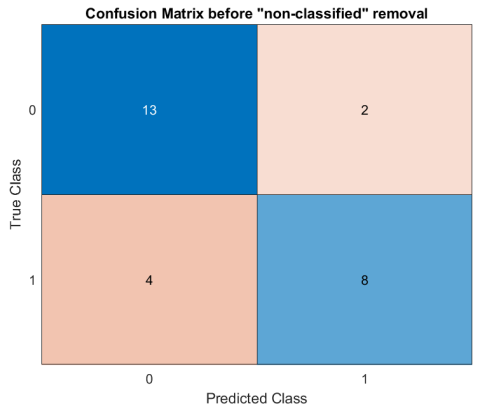

Figure 3.9: Confusion matrix for sustained vowel /a/ with $CPPS_{skewness}$ as selected feature before the removal of "non-classified"

Classified (FoC) are introduced; the definition of these two measures is offered in section 2.1.5. By excluding the elements belonging to the "non-classified" class, the new realistic confusion matrix and the new results of the evaluation metrics are presented (figures 3.11 and 3.12).

Figure 3.10: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for sustained vowel /a/ case after the removal of "non-classified"

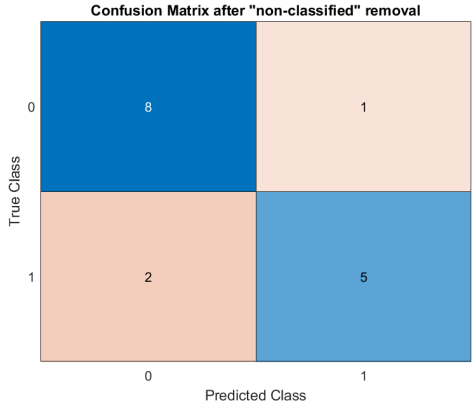

Figure 3.11: Confusion matrix for sustained vowel /a/ with $\text{CPPS}_{\text{skewness}}$ as selected feature after the removal of "non-classified"

Balanced speech task

The situation considered as the one with the best performance metrics between all three tasks is obtained with *gender*, $\text{CPPS}_{5,\text{prc}}$ and V/uV as selected features, returning an accuracy value equal to 92.3%. Being the number of selected features k equal three, the uncertainty formula (3.1) is applied taking into account four

HS (0) vs MS (1)	SUSTAINED VOWEL /a/ AFTER NON-CLASSIFIED REMOVAL					
Features	AUC_R	$Precision_R$	$Sensitivity_R$	$Specificity_R$	$Accuracy_R$	FoC
CPPS (skewness)	81,3%	83,3%	71,4%	88,9%	81,3%	59,3%

Figure 3.12: Classification performance obtained for sustained vowel /a/ with $CPPS_{skewness}$ as selected feature after the removal of "non-classified"

sensitivity coefficients and four covariances. Similar considerations to the ones reported in the sustained vowel /a/ task, are performed. The probabilities returned by the LR validated model are presented (figure 3.13). Since the performance of the classifier are high, the cases of wrong classification are only two; one subject belongs to Class 0, but is assigned by the LR classifier to Class 1 creating a FP and, in the other case, the subject belongs to Class 1 but it is classified as healthy (i.e., Class 0) so it becomes a FN.

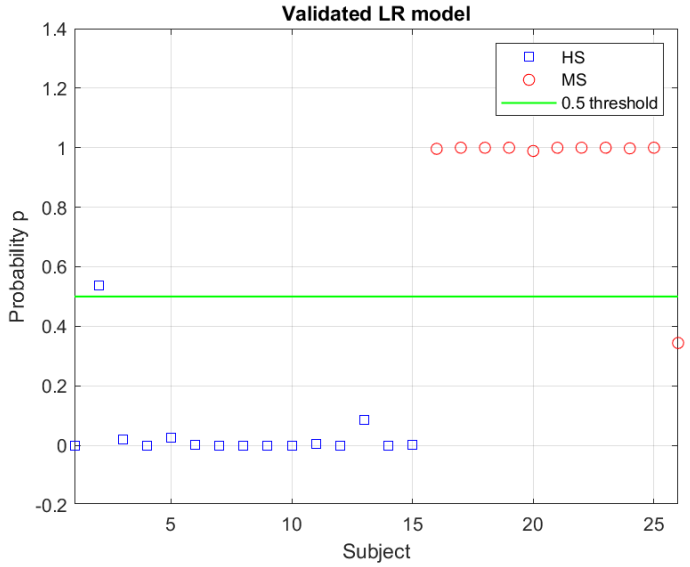

Figure 3.13: Probabilities returned by the LR validated model without the implementation of the expanded uncertainty for balanced speech task

The figure 3.14 shows the probabilities having obtained the expanded uncertainties. Being the uncertainty directly related to the sensitivity coefficients; these increase for probability values around 0.5, and decrease for probabilities near 0 and 1.

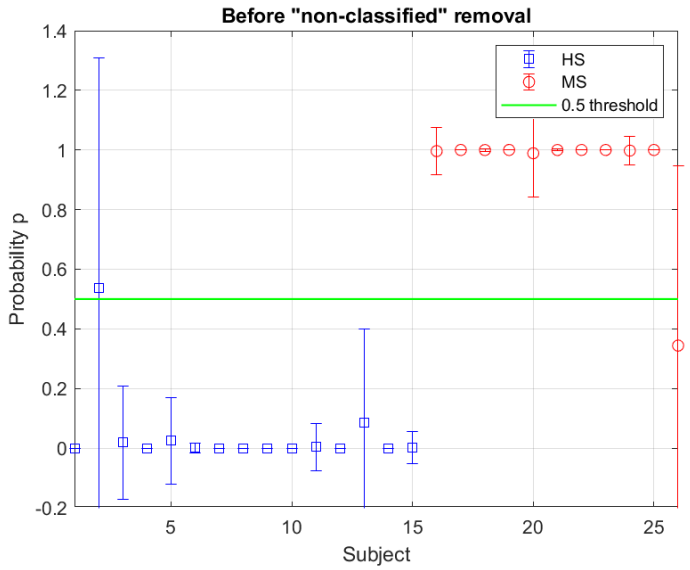

Figure 3.14: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for balanced speech task before the removal of "non-classified"

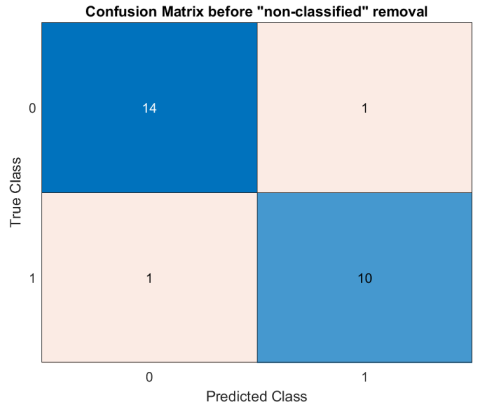

Figure 3.15: Confusion matrix for balanced speech task with *gender*, $\text{CPPS}_{5,\text{prc}}$ and V/uV as selected features before the removal of "non-classified"

After the removal of the "non-classified" subjects (represented in figure 3.16),

the total number of subjects is reduced from twenty-six to twenty-four. As it is possible to notice from the new confusion matrix (figure 3.17), by eliminating the two critical subjects from both classes, the number of FP and FN is remodeled equal to zero, which means that the classifier always predicts the correct label. In this case, the fraction of classified subjects is higher than in the previous task (FoC = 92.3%) and the other realistic metrics are exposed in figure 3.18.

Figure 3.16: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for balanced speech task after the removal of "non-classified"

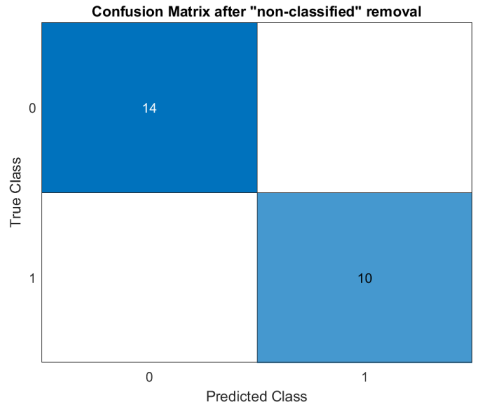

Figure 3.17: Confusion matrix for balanced speech task with *gender*, $\text{CPPS}_{5,\text{prc}}$ and V/uV as selected features after the removal of "non-classified"

HS (0) vs MS (1)	BALANCED SPEECH TASK AFTER NON-CLASSIFIED REMOVAL					
Features	AUC_R	$Precision_R$	$Sensitivity_R$	$Specificity_R$	$Accuracy_R$	FoC
GENDER, CPPS (5° perc), V/uV	100,0%	100,0%	100,0%	100,0%	100,0%	92,3%

Figure 3.18: Classification performance obtained for balanced speech task after the removal of "non-classified"

Free speech task

The implementation of the expanded uncertainty to the free speech case, is computed taking into account the feature combination consisting of HNR_{std}, $\text{CPPS}_{5,\text{prc}}$ and V/uV. Also, in this task the SE values and the covariance values of the β coefficients are extracted from the exported function in the Classification Learner App in Matlab and, the same steps in order to obtain the probabilities of the model are executed (as shown in figure 3.19).

As result having obtained the expanded uncertainty (figure 3.20), the vertical bars coupled to the probability values represent the confidence intervals.

In the case of four healthy subjects and three patients, the confidence interval includes the thick green line of the discrimination probability threshold at 0.5. These elements marked as "non-classified" are removed (in figure 3.22) and, then, the classification performance are updated taking into account this exclusion (see figures 3.23 and 3.24).

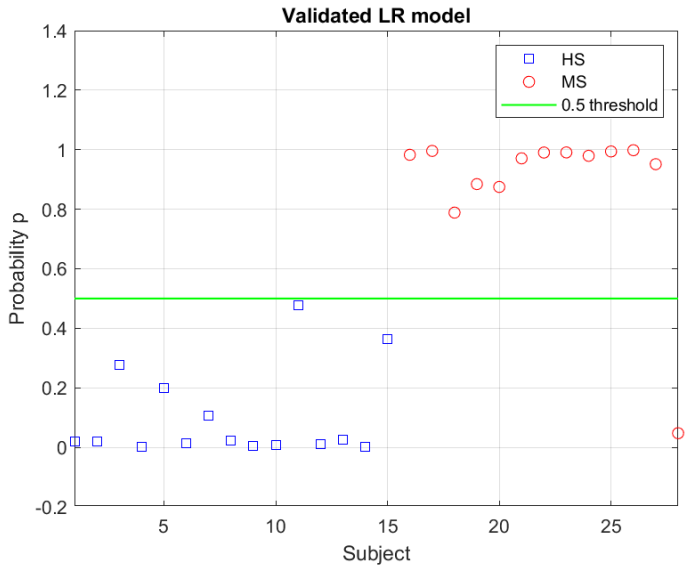

Figure 3.19: Probabilities returned by the LR validated model without the implementation of the expanded uncertainty for free speech task

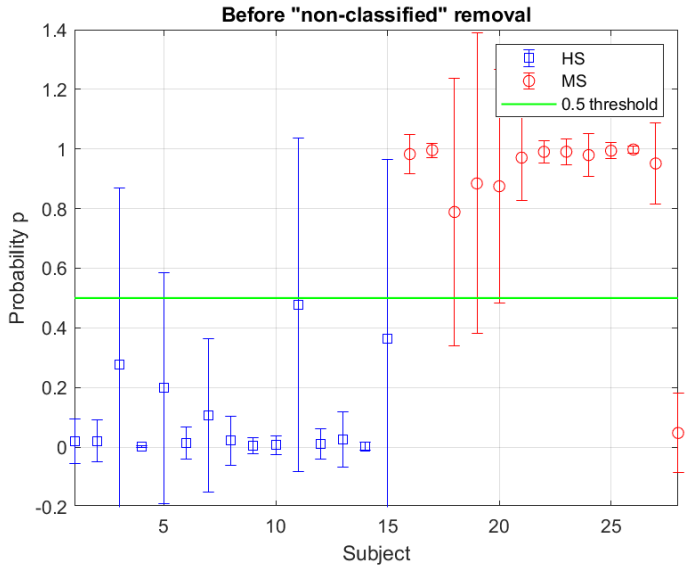

Figure 3.20: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for free speech task before the removal of "non-classified"

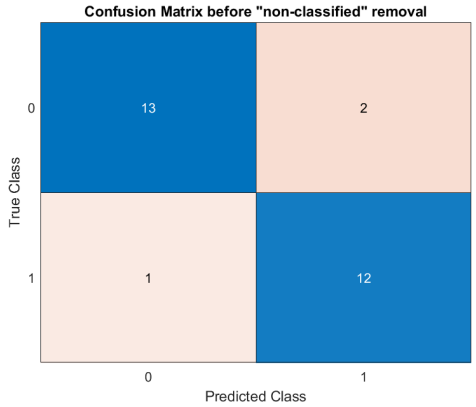

Figure 3.21: Confusion matrix for balanced speech task with HNR_{std}, $CPPS_{5,prc}$ and V/uV as selected features before the removal of "non-classified"

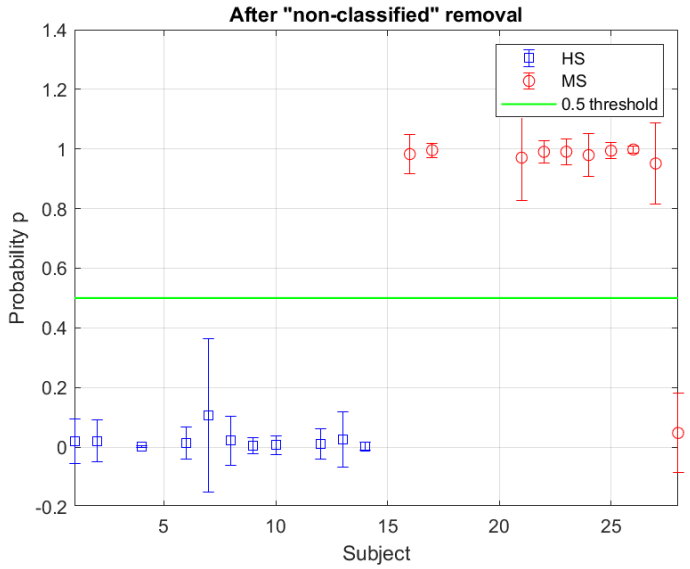

Figure 3.22: Probabilities returned by the LR validated model with the implementation of the expanded uncertainty for free speech task after the removal of "non-classified"

3.3 Validation in the use of VH

This section has the objective of justifying the use of the VH device as comparable tool with the in-air recording system and, at the same time, to characterize a certain group of subjects. VH has the ability to analyze vocal loading changes during a working day and to identify a person's risk of vocal dysfunctions; moreover this

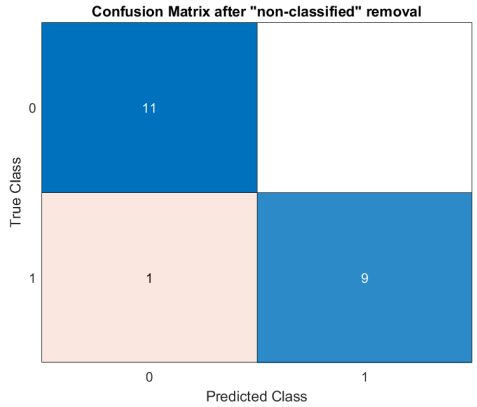

Figure 3.23: Confusion matrix for balanced speech task with $\mathrm{HNR}_{\mathrm{std}}$, $\mathrm{CPPS}_{5,\mathrm{prc}}$ and V/uV as selected features after the removal of "non-classified"

HS (0) vs MS (1)	FREE SPEECH TASK AFTER NON-CLASSIFIED REMOVAL					
Features	AUC_R	$Precision_R$	$Sensitivity_R$	$Specificity_R$	$Accuracy_R$	FoC
HNR (std), CPPS (5° perc), V/uV	95,0%	100,0%	90,0%	100,0%	95,2%	75,0%

Figure 3.24: Classification performance obtained for free speech task after the removal of "non-classified"

portable analyzer, can be considered as an aid for monitoring vocal health and vocal quality of subjects showing abnormal vocal cord vibration (thus, abnormal pitch and loudness). To investigate the equivalence between the contact microphone-based device and the in-air microphone (MIC), differences between the parameters extracted from MIC (obtained in section 2.1.2) and the ones stored inside the DAP unit of VH are computed. These differences Δ (as already exposed in section 2.2.1), refer to the application of the Eq. (2.22) for all the parameters considered in this part. Recalling that, the long-term monitoring is performed with VH only, this delta analysis is executed for short-term assessment (i.e., sustained vowel /a/, balanced and free speech task), common to both the microphones. It is valuable to emphasize that, the total number of subjects do not change during the evolution of this thesis work; the involved subjects are thirty-two (sixteen healthy subject and sixteen MS patients). Whether in a certain investigation the total number of subject is different, this can be attributed to the fact that, for some subjects (both HS and MS), the files needed to compute this comparison are not available or exhibited non-significant values (e.g., CPPS value equal zero) dissuading, in this case, their inclusion in the investigation.

Sustained vowel /a/

Considering the repetitions of vowel /a/ task, the parameters in common with the in-air microphone and VH are: local jitter (%), local shimmer (%), $\text{CPPS}_{\text{median}}$ (dB) and CPPS_{std} (dB) (listed in figure 2.6 in section 2.2.1). From this investigation four figures are derived as output, being four in number the parameters on which the comparison between MIC and VH is carried out. The subjects are twenty-three, of which the first eleven are MS patients (indicated as red elements), while the healthy subjects are twelve (associated with blue color). Delta values for each subject, are indicative of how comparable (or at least congruent) the parameters extracted from the MIC and the ones processed by VH device are to each other. The differences, obtained between the two microphones, are not negligible, but delta values equal zero are not expected, since the two input signals for each subject are not the same. For the VH, the input signal consists in the vibration (mechanical signal) induced by the vocal folds at the neck, the latter considered as a low-pass filter; while an in-air pressure signal, modulated by the vocal tract, is the signal acquired with the the in-air microphone; in addition, the two devices present a different measurement chain (i.e., the bandwidth). Since the devices have different characteristics, the use of VH device (as an aid for assessing vocal health) requires the definition of specific cut-off values for the extracted parameters. With more detail, according to the following rules expressed in [28], voice recordings are classified as healthy, pathological or "not reliable" in dependence on the values of the aforementioned parameters.

- A local jitter value $< 0.31\%$ identifies an healthy voice, $> 0.43\%$ a pathological voice and if is comprised between $(0.31 - 0.43)\%$, then it is considered not reliable.

- If local shimmer value is $< 2.37\%$ identifies an healthy voice, $> 2.55\%$ a pathological voice and if is in the range $(2.37 - 2.55)\%$, then it is considered not reliable.

- If $\text{CPPS}_{\text{median}}$ value is > 19.7 dB identifies an healthy voice, < 18.0 dB a pathological voice and if is comprised between $(18.0 - 19.7)$ dB, then it is considered not reliable.

- A CPPS_{std} value < 0.9 dB identifies an healthy voice, > 1.3 dB a pathological voice and if is in the range $(0.9 - 1.3)$ dB, then it is considered not reliable.

Additionally, VH device is preferable both for convenience in conducting acquisitions (i.e., the subject is free to move, without the need to worry about the distance between the mouth and the microphone) and, for its insensitivity to other possible sound sources in the environment. It is particularly meaningful to examine the

results of deltas relative to the amplitude and period stability parameters (i.e., local shimmer and local jitter respectively), that are available only in this task. In the case of local shimmer measure (figure 3.25) significant differences, between the parameters extracted from MIC and VH, are founded. Having a positive (above zero) mean value of delta shimmer, means that the measure of shimmer in air is higher than the one extracted from VH; this result is reliable, since the stability in amplitude of MIC can be worse in respect to the contact microphone-based device, being the latter a more stable analyzer, less affected by possible sources of noise in the environment. The black line, representing the mean value of delta shimmer, (equal to 4.38%) is significantly higher if compared to the cut-off values already mentioned (for local shimmer) and, it provides information on the average difference between MIC and VH between all the subjects. The delta shimmer value of the subject 17 (in figure 3.25) is highly negative if compared to the mean value of delta shimmer (i.e., the black horizontal line); this result, suggest that an error may have occurred during the acquisition of that subject with the VH device (for e.g. a displacement of the collar, which should not be moved for the entire duration of the recording) and this case, can be considered as an outlier. Generally,

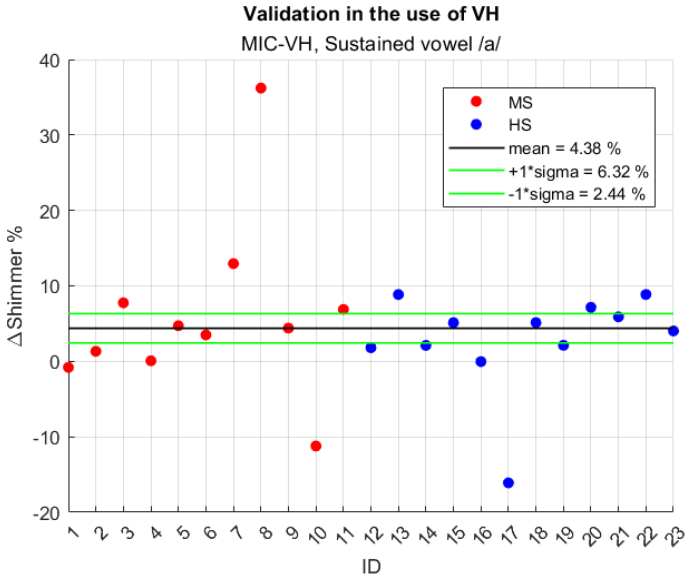

Figure 3.25: Results of delta shimmer values for each subject in the case of sustained vowel /a/ task

a worsening in the value (i.e., an higher value) of shimmer in air is expected, which represents the stability in amplitude, compared to a less evident delta jitter value (and thus, a value more near zero), which indicates that the in-period stability (i.e. jitter) is less affected by background noise sources (as reported in figure 3.26).

This result is corroborated by looking at the mean value of local jitter (the black line), being just above the unreliable range (relative to the mentioned cut-off values), which allows to classify a subject as healthy or pathological. In the case of

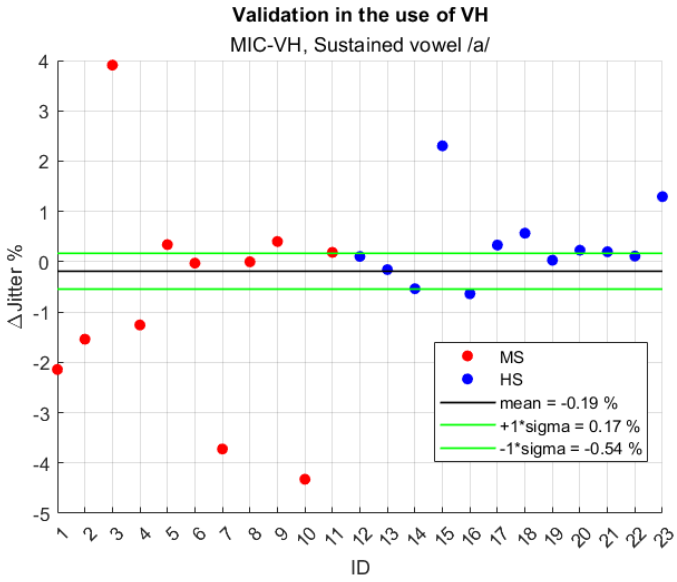

Figure 3.26: Results of delta jitter values for each subject in the case of sustained vowel /a/ task

$CPPS_{std}$ (as for local jitter), no significant delta values are observed. The mean value of delta $CPPS_{std}$ is negative (figure 3.27), which means that the $CPPS_{std}$ value measured in air is slightly lower than the one provided by VH. This results is reliable, since the signal recorded by the microphone in air can be compromised by noise sources, therefore, it is strongly related to the background noise level of the environment in which the acquisitions take place (e.g. in-clinic). In addition, the measurement chains' bandwidth of the two microphones are different: while the contact microphone-based device has a frequency content of approximately 3.5 kHz, for the in-air microphone case, this is 10 kHz. Referring to the results exhibited in [10], $CPPS_{5,prc}$ and $CPPS_{std}$ change when they are estimated from devices characterized by different bandwidths. Since the devices have different characteristics, it is important to assert that a mean value of delta equal to zero cannot be expected.
For the parameters local jitter, $CPPS_{median}$ and $CPPS_{std}$ the validation can be considered passed, while for the others, such as local shimmer, significant differences in terms of Δ value is noted. An attempt to improve the mean value of delta shimmer, is applied by removing elements showing abnormal behavior in respect to the delta mean value (e.g. subject 17) and, among the subjects that remain, a

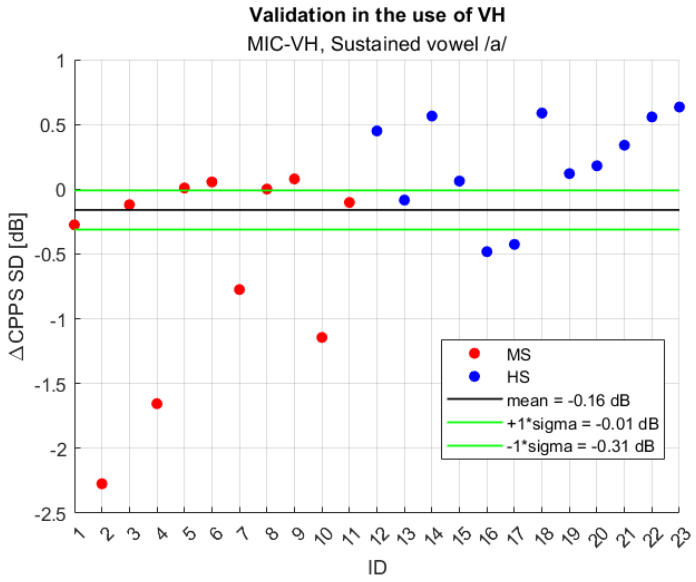

Figure 3.27: Results of delta CPPS_{std} values for each subject in the case of sustained vowel /a/ task

new mean value of delta shimmer is calculated. Outliers are selected as subjects exhibiting a value of delta shimmer exceeding three times the confidence interval set to $\pm 1 \cdot \sigma$ (i.e., $\pm 3 \cdot \sigma$, indicated with yellow lines in figure 3.28). As result, the mean value of delta shimmer (indicated with magenta line) remains high (equal to 4.15%), especially if compared with the unreliable range applied to this parameter (i.e., $(2.37 - 2.55)\%$); even removing the outliers from the calculation of the mean value of delta shimmer, the validation of the VH for this parameter, can not be considered achieved.

Balanced/free speech task

In the balanced and free speech case, the examined comparisons are available for all nine descriptive statistics of fundamental frequency f_0 (Hz) and CPPS (dB) (as displayed in figure 2.7 in section 2.2.1). In the balanced speech task, the total number of subjects involved is twenty-five (1-11 MS patients as red elements and 12-25 healthy subjects as blue ones), while in the case of free speech they become twenty-seven (which are 1-13 MS and 14-27 HS); these differences in the number of subjects, derive from the availability (for each subject) of both, the file returned by VH device and the .wav acquisition from the in-air microphone. Although, the sustained vowel /a/ is considered an unnatural vocalization if compared with the reading and free speech, for these tasks congruent results are obtained regarding the validation of VH device. The mean value of delta $\text{CPPS}_{5,\text{prc}}$ is not zero (as

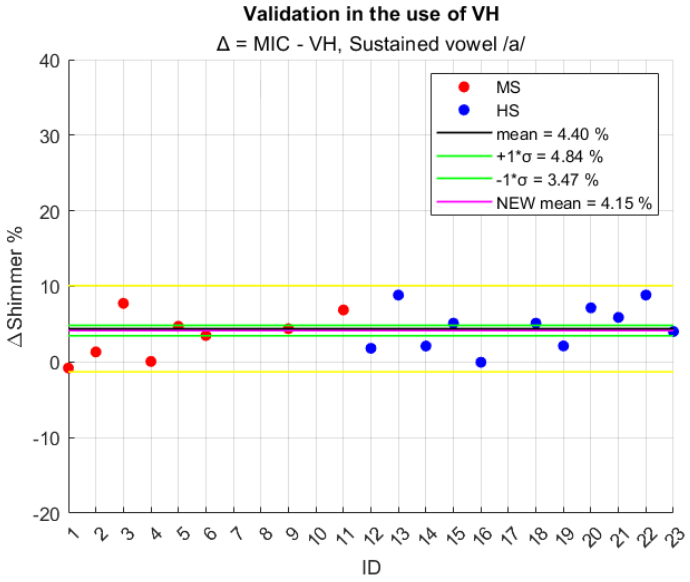

Figure 3.28: Results of delta shimmer values for each subject after the outliers removal in the case of sustained vowel /a/ task

expected), since the bandwidth of the measurement chains of the two microphones are different, but it is negative (figure 3.29); this could derive from the presence of noise sources in the in-air microphone acquisition, leading to a worsening in the value (i.e., a lower value) of CPPS in air.

3.4 Assessment of vocal fatigue

The method proposed to assess fatigue is presented in section 2.2.3 and, in this case, only data acquired with VH device are considered. Differences δ between the parameters extracted from the long-term and the correspondent short-term monitoring (i.e., the balanced and free speech task) are calculated, the last considered as a sort of "baseline" (i.e., the parameters during the first instants of the evaluation). This comparison is carried out considering the parameters Sound Pressure Level (dB), fundamental frequency f_0 (Hz), CPPS (dB), and Background Noise Level (90° percentile) in dBA (as shown in figures 2.10 and 2.11). It is expected that, while performing the baseline, no difference between the two classes (HS and MS) is noted, as the time proceeds, a distinction between healthy subjects and patients will emerge, being fatigue one of the most obvious and debilitating symptoms of Multiple Sclerosis. To perform this investigation, a mean value of each parameter is extracted for both the acquisitions (long-term and short-term monitoring), for each subject; then, delta values are calculated, but no significant differences in terms

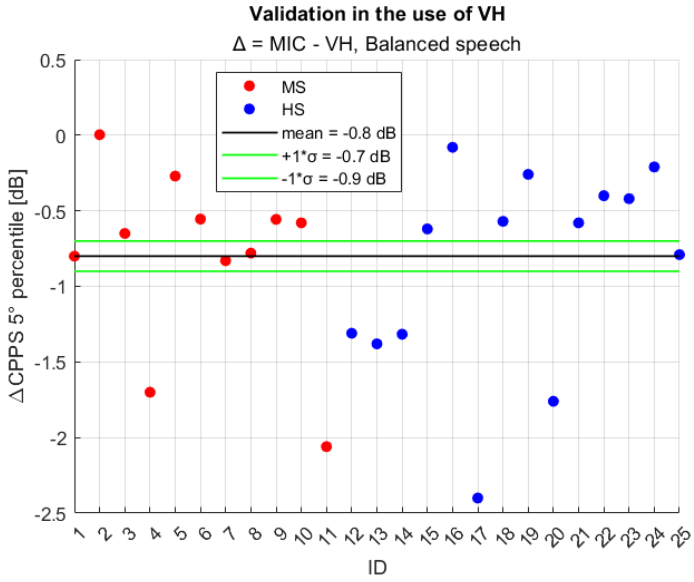

Figure 3.29: Results of delta $\text{CPPS}_{5,\text{prc}}$ values for each subject in the case of balanced speech task

of fatigue experienced between the two classes is noted. Considering the SPL_{mean} parameter, both HS and MS show a positive mean value (as reported in figure 3.30), which means that there is an increase in speech intensity in long-term evaluation, if compared to baseline (the latter being an in-clinic monitoring performed under comfortable conditions). Actually, this result is consistent with the discussion exhibited in [29], where an increase in fundamental frequency and sound pressure level after long periods of voice usage is reported (whether for professional use or not). The red band and the blue one (the confidence intervals set to $\pm 1^{*}\sigma$ for MS and HS respectively) are not well separated and, as a consequence, it is not possible to distinguish a significantly different behavior of MS patients compared to healthy subjects. In addition, it is reported that, two strategies during vocalization are used in order to increase vocal intensity; these are, using more energy (i.e., air) or raising the fundamental frequency through compensatory strategies (e.g., using different muscles of vocal tract with reference to the Appendix A.1); the latter, being less advantageous, but more typical of subjects developing dysphonia or fatigue. An effect of this phenomenon is reported (in figure 3.31), where the MS class presents a mean value of $f_{0,\text{mean}}$ slightly higher in respect to HS and a gap between the two confidence intervals is noted.
It is important to express the fact that, the results obtained for different parameters depend on the environment in which the acquisitions are carried out, since a strong correlation of the data with the background noise level is expected. The

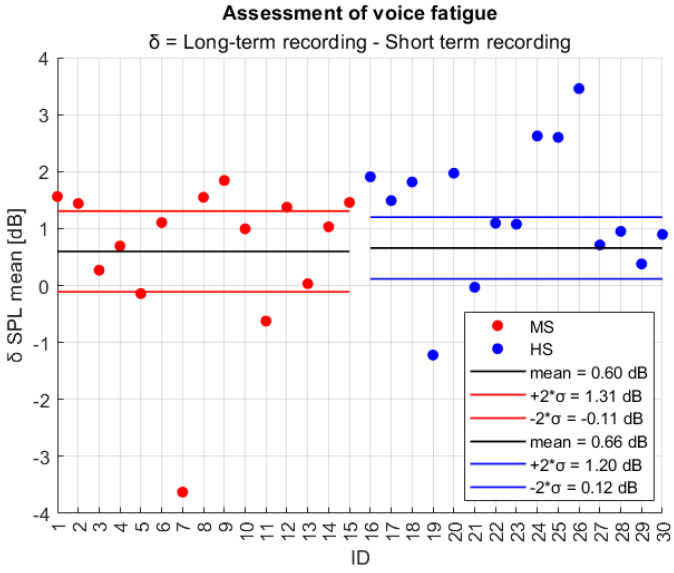

Figure 3.30: Results of delta SPL_{mean} values in the comparison between long-term and short-term evaluation

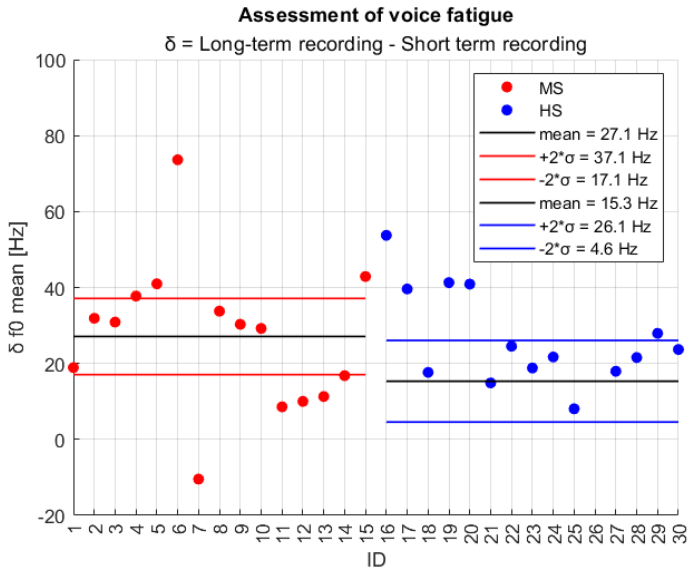

Figure 3.31: Results of delta $f_{0,\text{mean}}$ values in the comparison between long-term and short-term evaluation

parameters acquired with VH for the long-term evaluation, are visualized over time with reference to the background noise level (as shown in figures 3.32 and 3.33).

The red colour is associated with the MS class, while the blue one with HS group. It is clear, that, a more complete analysis of the trends of these parameters can be done only with the use of diaries, reporting both the environment where the acquisitions take place and the activity performed by the subject. To conduct

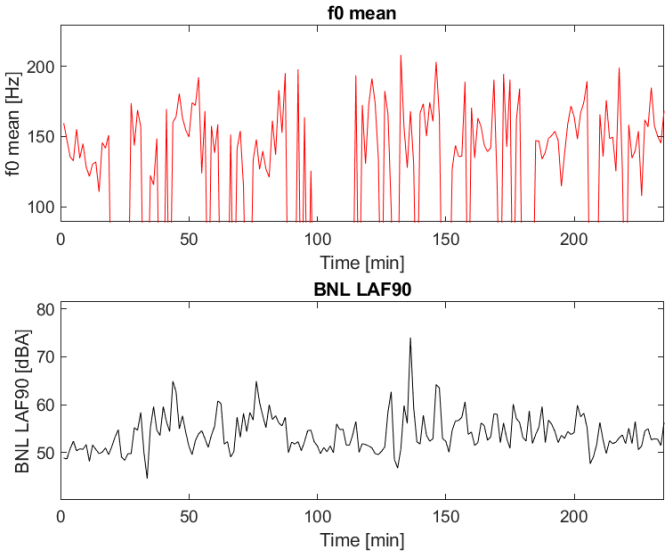

Figure 3.32: Representation of $f_{0,\text{mean}}$ parameter over time with reference to the BNL_{LAF90} value present in the environment in the case of one patient

this investigation, a removal of unvoiced frames is carried out for all parameters, except for BNL_{LAF90} parameter, being its value significant regardless of whether the subjects is talking or not. Observing the figures, it is possible to notice that where the parameter becomes zero, it means that, in that frame, the subject is not speaking.
It is interesting to observe the correlations between the parameters SPL_{mean}-BNL_{LAF90}, $f_{0,\text{mean}}$-BNL_{LAF90} and SPL_{mean}-$f_{0,\text{mean}}$. The regression line (which fits the correlation between these parameters) and the correlation coefficient R^2 (which expresses the goodness-of-fit of the model) are calculated with the idea of performing an inter-class and an intra-class analysis. If the environment is characterized by a loud background noise level, then one strategy that subjects usually perform in order to prevail, is to raise their speech intensity of voice (i.e., the SPL parameter). This effect, is represented in the map SPL_{mean}-BNL_{LAF90} as an increase in the regression line. Additionally, considering that MS patients suffer from hypophonia, it is expected that they are unable (or less able than healthy people) to increase their speech intensity level as noise increases. Observing the subjects (both MS and HS) taken individually, it is not possible to notice any significant difference in

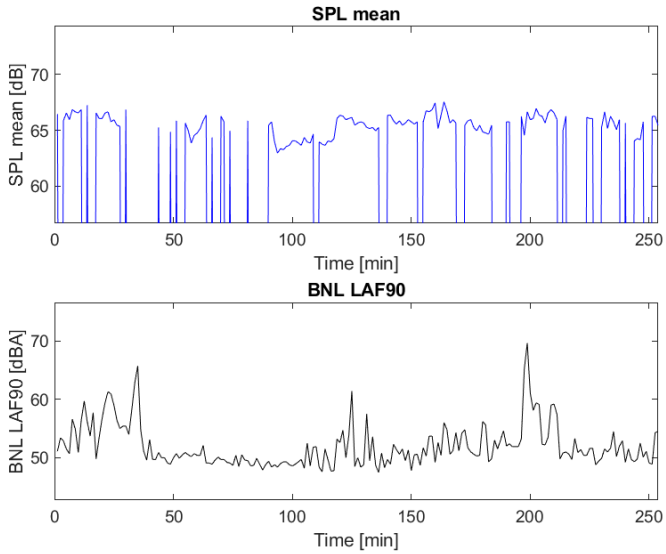

Figure 3.33: Representation of SPL_{mean} parameter over time with reference to the BNL_{LAF90} value present in the environment in the case of one healthy subject

respect to the values of slope of the regression model; in fact, the angular coefficient of the regression line in the case of MS patients (figure 3.34) does not assume lower values if compared to the one of healthy subjects (figure 3.35).

However, with reference to the same map (i.e., SPL_{mean}-BNL_{LAF90}), considering no longer the individual subject, but the entire MS/HS class, then the results are in accordance with what expected. The MS patients (in red) show a negative slope for the regression model in the speech intensity-noise level correlation, in contrast to the healthy subjects (in blue) that present a positive angular coefficient; this means that, if there is an increase in the BNL_{LAF90}, generally the MS patients decrease their speech intensity over time (as shown in figures 3.36 and 3.37).

Observing the SPL_{mean}-$f_{0,\text{mean}}$ correlation in the case of healthy subjects, this map (figure 3.38) has a definite pattern; if HS increase the speech intensity, they also increase their fundamental frequency value and in the map, this results in an asymmetrical shape characterized by a peak in the upper right corner.
In order to observe fatigue in the representation of the SPL_{mean} parameter over time (i.e., a decrease in the value of speech intensity over time), it is important to exclude the effect of the background noise level; since an increase in the noise level, can lead the subject to increase (even unintentionally) the intensity of the voice and thus, to hide the effect of fatigue. To perform this analysis, the angular coefficient of the regression line modeling the SPL_{mean}-BNL_{LAF90} correlation is used (figure 3.34). In this case, the MS patient is able to increase the speech

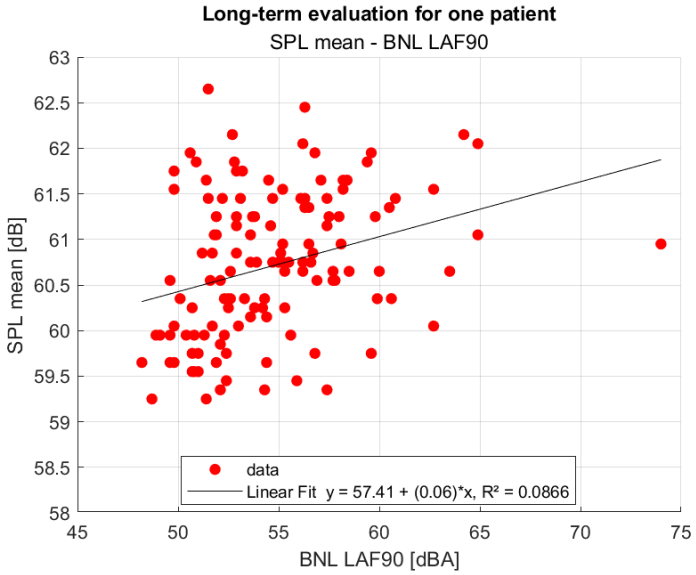

Figure 3.34: Representation of SPL_{mean}-BNL_{LAF90} correlation in the case of one patient

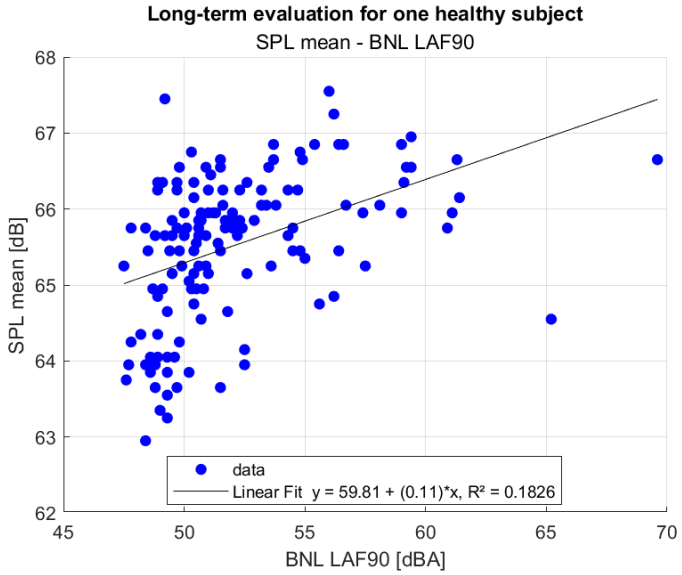

Figure 3.35: Representation of SPL_{mean}-BNL_{LAF90} correlation in the case of an healthy subject

intensity level in order to cope with an increase of BNL_{LAF90} parameter in the environment (since the slope of the regression model is positive and equal to 0.06

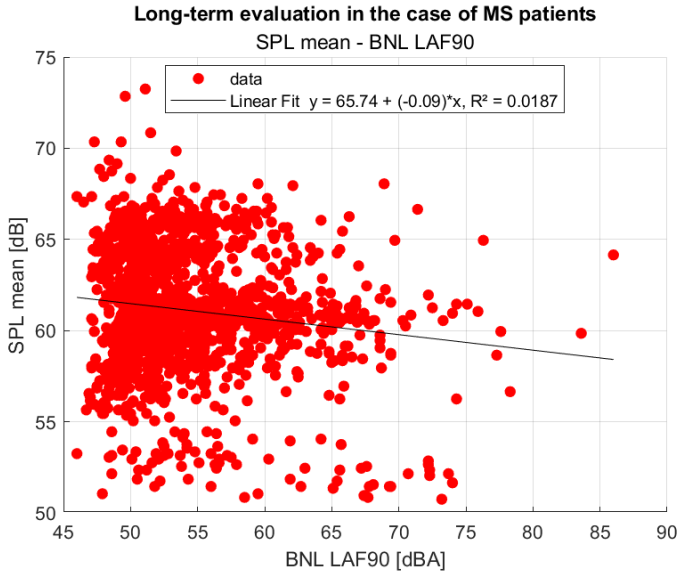

Figure 3.36: Representation of SPL_{mean}-BNL_{LAF90} correlation for MS class

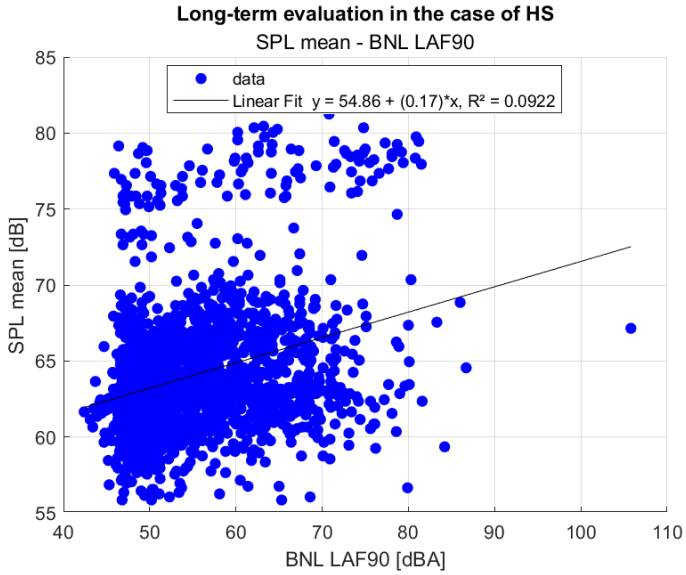

Figure 3.37: Representation of SPL_{mean}-BNL_{LAF90} correlation for HS class

dB/dBA). The initial SPL_{mean} value before performing the BNL_{LAF90} correction (represented in red) is equal to 60.31 dB (i.e., the intercept at time zero) and it increases of approximately 0.005 dB/min over time (as shown in figure 3.39). This compensation is executed by operating for each frame a difference between the

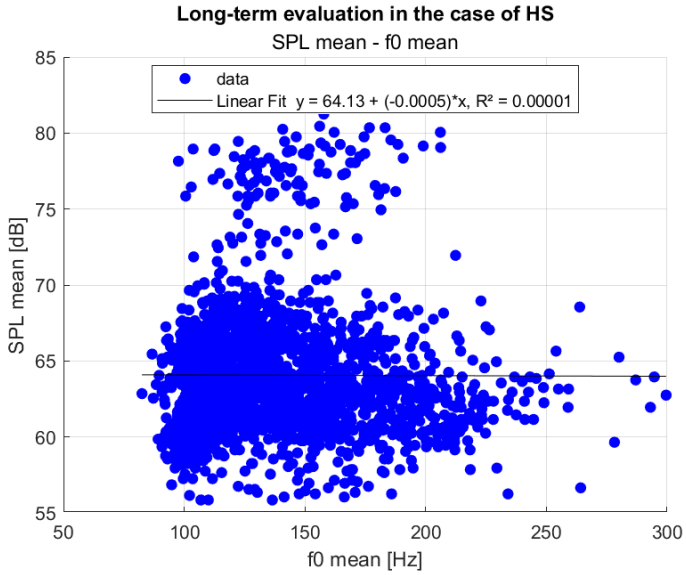

Figure 3.38: Representation of SPL_{mean}-$f_{0,\text{mean}}$ correlation for HS class

original SPL_{mean} value and the BNL_{LAF90} parameter, multiplied by the angular coefficient of the SPL_{mean}-BNL_{LAF90} regression line (which, as already mentioned, is equal to 0.06 dB/dBA). The "corrected" SPL_{mean} value over time (represented in

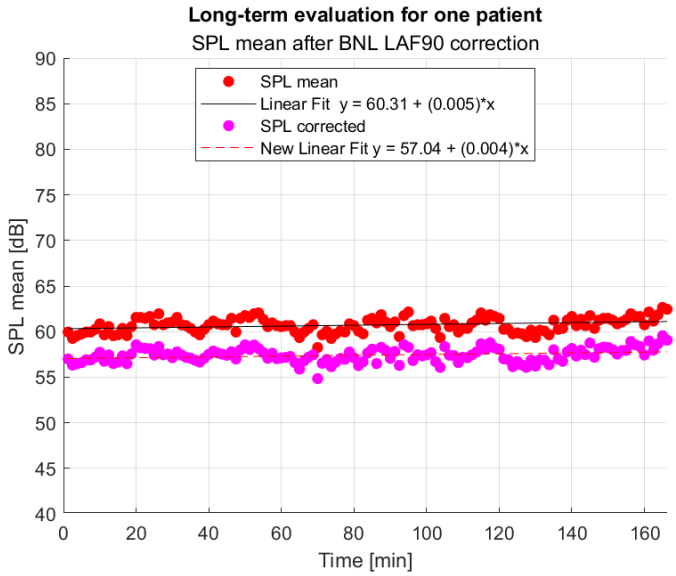

Figure 3.39: Representation of SPL_{mean} compensation in respect to the BNL_{LAF90} over time for one patient

magenta) indicates that, in absence of noise in the environment, the subject starts speaking at 57.04 dB and the speech intensity increases of approximately 0.004 dB/min over time. Also in this case, the subject does not show fatigue since the value of the angular coefficient is positive. To observe differences between HS and MS class, the value of the angular coefficient of the regression line (modeling the parameter SPL_{mean} corrected over time) for each subject is extracted and plotted on a graph; on the y-axis the value of the slope in dB/min is reported, while on the x-axis there are the MS patient (in red) and the healthy subjects (in blue). An higher value of slope is expected for healthy subjects in respect to MS, since the patients should show an higher fatigue. In this case, a difference between the two classes is not observed (figure 3.40). It is possible that, the mean value of the slope of the regression line for the healthy subjects (equal to 0.0032 dB/min in figure 3.40) is lower in respect to MS, since the time interval of the long-term recordings for the healthy group is longer and therefore, they show more fatigue. In addition, this hypothesis is confirmed by observing that, the MS group exhibits a time interval duration for the long-term monitoring, ranging from a minimum of 95 to a maximum of 247 minutes, while in the HS case, this ranges from 202 to 311 minutes. With the aim of exposing more comparable results in terms of fatigue in future acquisitions, it is important that all subjects (included in the data-set) display approximately equal time intervals duration for the long-term monitoring.

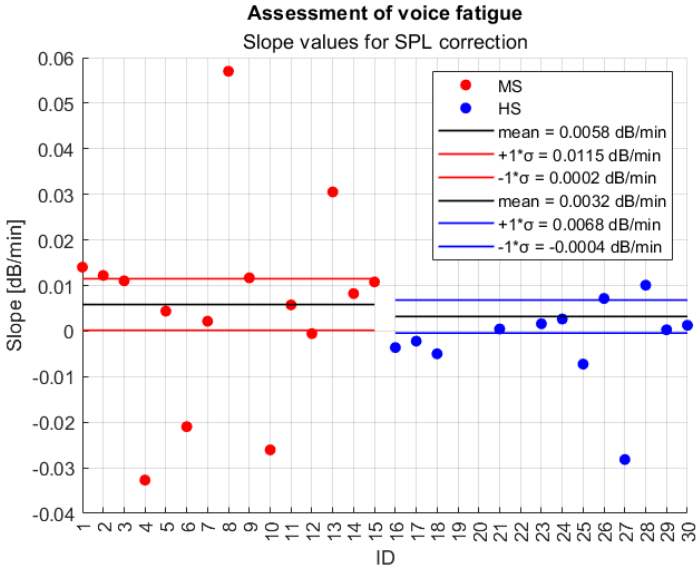

Figure 3.40: Representation of the extracted slope values (dB/min) of the regression line modelling the SPL_{mean} parameter compensated in respect to the BNL_{LAF90} over time for each subject

Observing the results obtained by extracting the SPL_{mean} parameter over time, it is possible to notice that, MS patients exhibit lower variability in respect to healthy subjects. In particular, the elements (i.e., the SPL_{mean} values measured by VH in each 75-s frame) in the case of MS patients, seem to be more centered on the regression model fitting the data (as shown for MS and HS in figures 3.39 and 3.41 respectively). From these results, it is expected that MS patients show lower ability

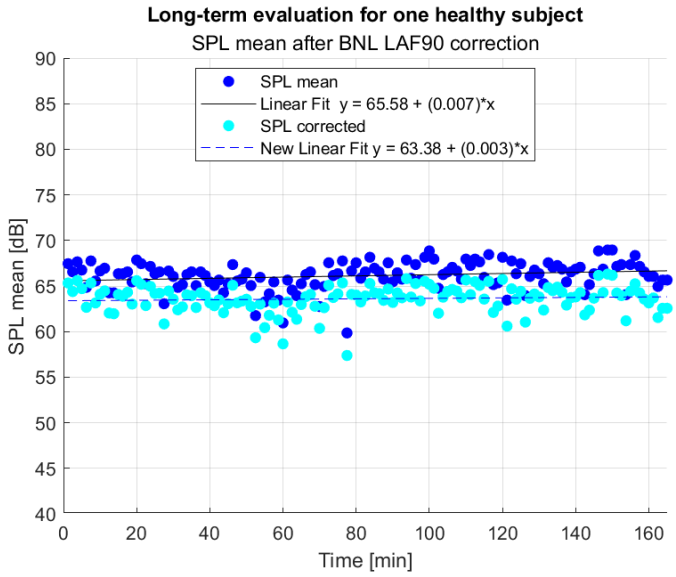

Figure 3.41: Representation of SPL_{mean} compensation in respect to the BNL_{LAF90} over time for an healthy subject

to vary both speech intensity level (SPL) and fundamental frequency (f_0), and furthermore, exhibit a significantly lower range of variability than healthy subjects. To assess the variability of speech intensity level over time without BNL_{LAF90} correction (since it is not necessary to consider the parameter SPL_{mean} corrected over time), the value of the SPL_{std} parameter for each subject is extracted and plotted on a graph. The results (shown in figure 3.42) show that MS subjects have a mean value of SPL_{std} equal to 2.53 dB, compared to healthy subjects showing a mean value of 2.91 dB; this means that MS subjects present slightly reduced variability in intensity if compared to healthy subjects. Similarly, an analysis on the ability of different subjects in varying their fundamental frequency f_0 is conducted. The procedure to extract $f_{0,\text{std}}$ parameter is the same, as already performed for SPL_{std}. The results (expressed in figure 3.43) show that the blue band, representing the ability of healthy subject in varying their frequency f_0, is slightly higher (reporting a mean value of $f_{0,\text{std}}$ equal to 34.35 Hz). However, limitations in this experiment can be overcome through an increase in the data-set,

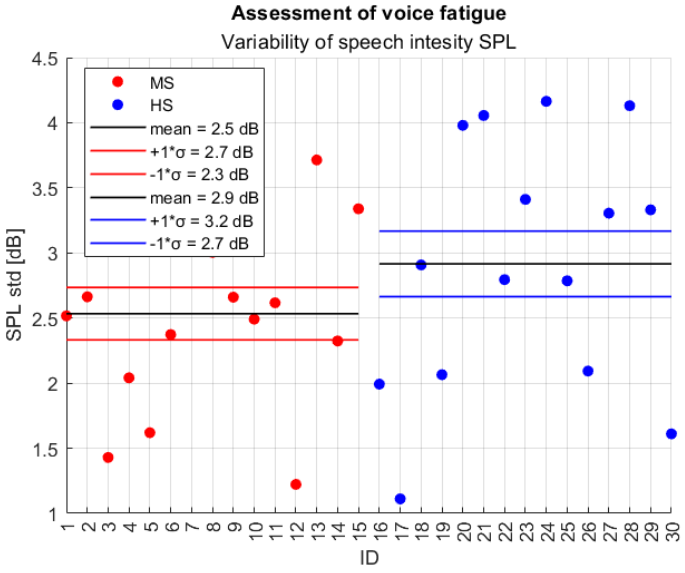

Figure 3.42: Representation of the mean value of SPL_{std} for each subject for the long-term monitoring

so as to narrow these bands and to observe significant differences between the classes.

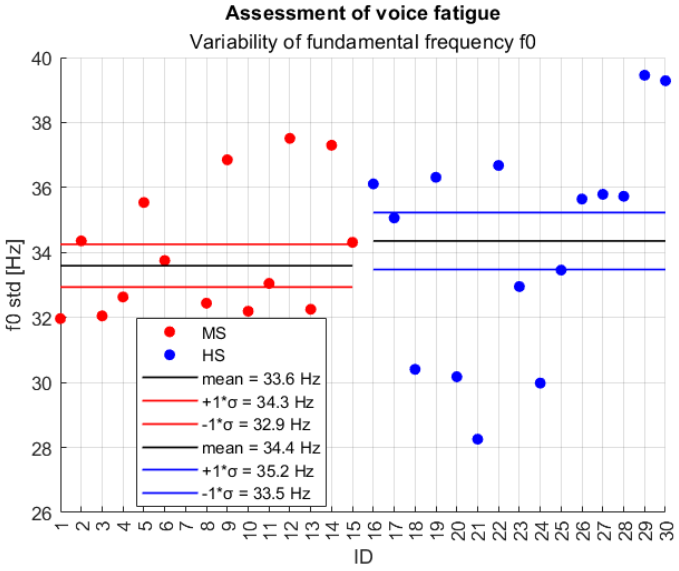

Figure 3.43: Representation of the mean value of $f_{0,\text{std}}$ for each subject for the long-term monitoring

3.5 Vocal doses

Excessive vibration of vocal fold tissues, due to loud or prolonged vocalization, is assumed to contribute to the development of voice disorders and present a significant health concerns. The five vocal doses measures the vocal load and can be used for studying the effects of vocal fold tissue exposure to vibration, which is experienced by subjects in long-term assessment. Having defined the vocal doses (in section 2.2.4), their accumulation for each subject is demonstrated for the entire duration of the long-term acquisition (i.e., the total measurement time, t_m); in addition, each dose is referred to the parameter k_v, defined as the voicing unit step function. All the doses tend to increase when the voicing function is equal to 1 (voiced frame) and, remain constant when k_v assumes the value of 0 (unvoiced frame). The dose measures allow different vocalizations to be compared and, can be used in studying different factors potentially harmful for the vocal folds. The examples of accumulation of all five vocal doses are shown for the case of one MS patient (in red). The time dose (figure 3.44), although sensitive to neither frequency nor loudness, can be used for quantifying the duration of voicing percentages among various vocal activities or occupations. The cycle dose (figure 3.45) is found to

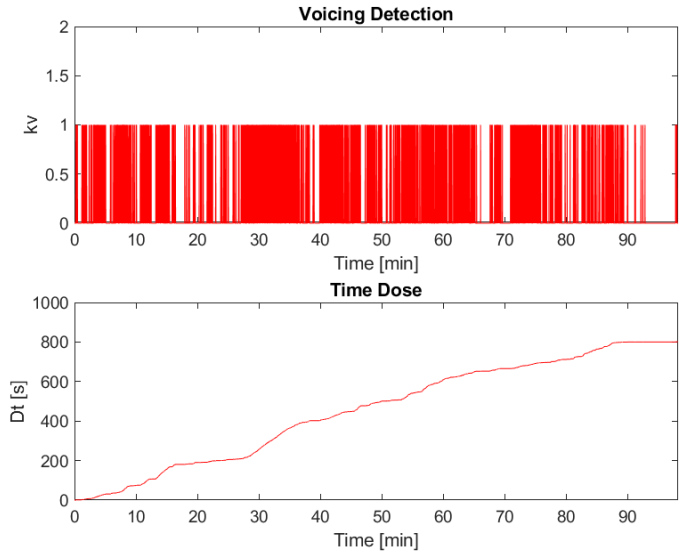

Figure 3.44: Accumulation of the time dose and the correspondent voicing unit step function in the case of one patient

contribute to the correlation of the larger number of cycles with the larger number of vocal complaints, deriving from the potentially harmful effect of collisions of the vocal folds. The distance dose measures the total distance accumulated by the

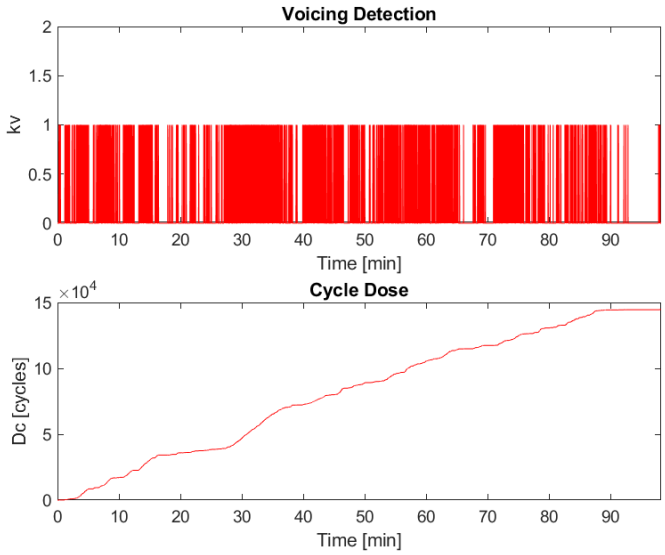

Figure 3.45: Accumulation of the cycle dose and the correspondent voicing unit step function in the case of one patient

vocal folds during vibration; two examples (figures 3.46 and 3.47) show the cases of distance dose (D_d) for a male patient in red and an healthy male subject in blue. D_d in the MS case reaches a value just above 350 m after a total time interval for long-term monitoring of approximately 98 minutes; while, in the HS case, it reaches a value above 400 m in a total measurement time of about 223 minutes.

The energy dissipation dose calculates the total amount of heat generated in a unit volume of vocal fold tissues; however the amount of heat generated in the vocal folds is smaller than expected [23] (as presented in figure 3.48). The radiated energy dose can be used for studying the efficiency of the voice production, relating the energy radiated out the mouth to the total energy dissipated in the vocal folds; however this is only a crude estimate (figure 3.49), since a proper quantification of the total dissipated energy would require the knowledge of the length, the thickness and the depth of the vocal fold vibration for each subject.
From here, it is possible to understand that the extracted doses exhibit orders of magnitude comparable with values found in literature [24]; however, an important limitation is presented, which is the total measurement time for the long-term monitoring, since it is not fixed, but varies for each subject (as it is possible to observe from the time axis in figures 3.46 and 3.47). All subjects under analysis performed long-term monitoring ranging from 95 minutes up to 326 minutes. Unless during the acquisition, vocal effort is significant among the other subjects, then, there is the problem of being able to compare the five vocal doses. To compute

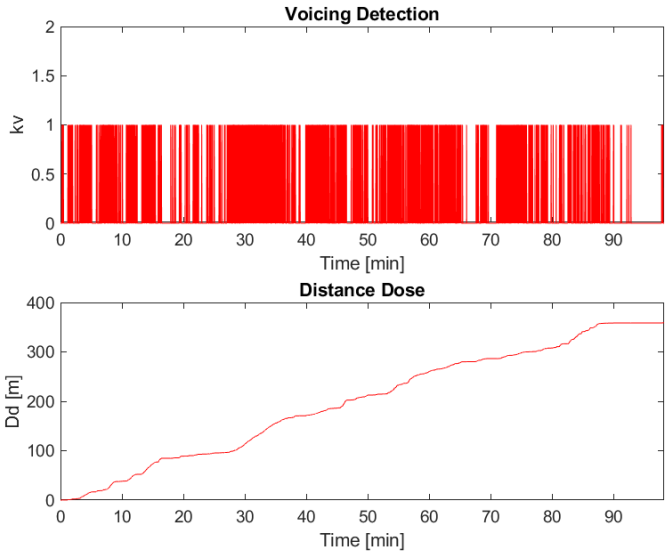

Figure 3.46: Accumulation of the distance dose and the correspondent voicing unit step function in the case of one patient

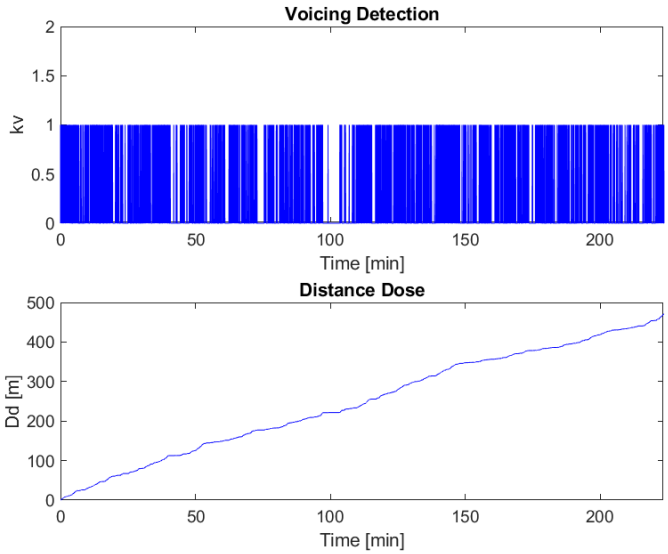

Figure 3.47: Accumulation of the distance dose and the correspondent voicing unit step function in the case of an healthy subject

an assessment at consistent times, a minimum time interval duration common to all subject is considered. In this case, the minimum time interval duration chosen

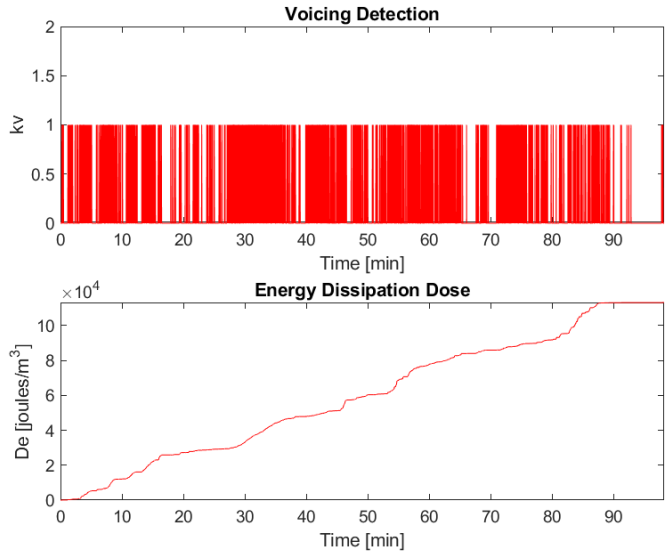

Figure 3.48: Accumulation of the energy dissipation dose and the correspondent voicing unit step function in the case of one patient

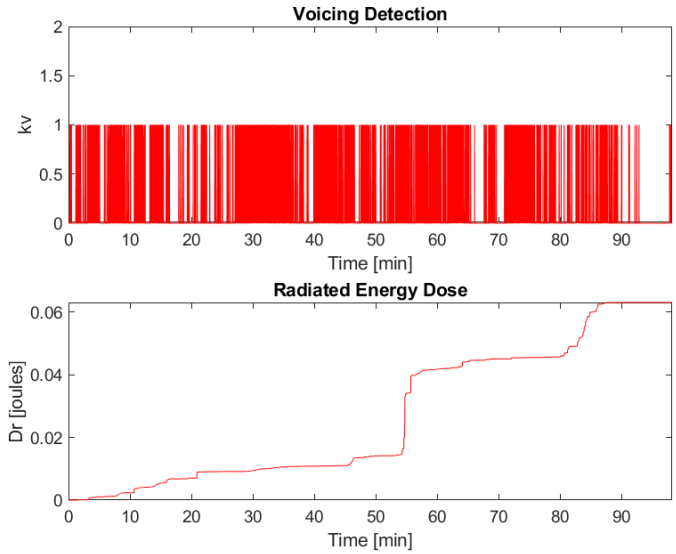

Figure 3.49: Accumulation of the radiated energy dose and the correspondent voicing unit step function in the case of one patient

is 95 minutes. For each subject, the five vocal doses at minute 95 are extracted and represented. Observing the results in the case of the time dose as a function

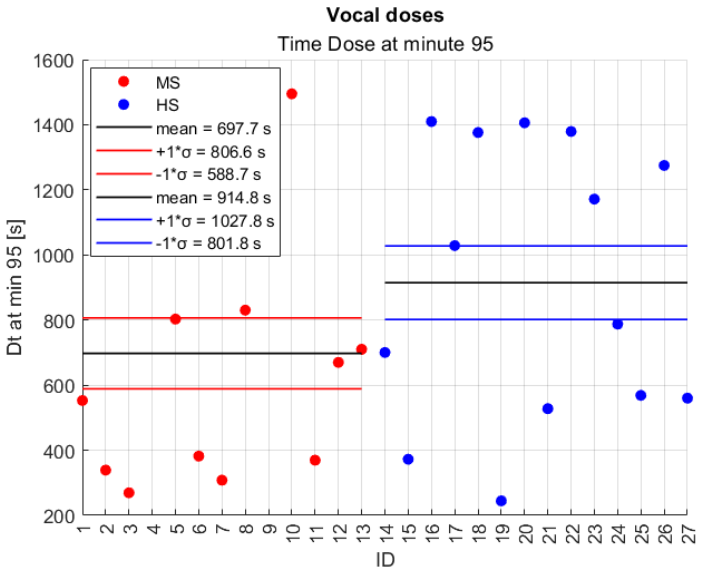

Figure 3.50: Time dose values at the minimum time interval of 95 minutes for each subject under analysis

of the subjects (figure 3.50), similar D_t values between HS and MS are founded. Regarding a minimum time interval of 95 minutes, the time dose for different subjects ranges from 244 to 1815 seconds; healthy and pathological subjects are mixed and, no conclusions as to whether the patients speak less than HS can be performed; thus, no differences, in this case, are noticed and the time dose parameter is not significant in distinguishing the two classes.
Considering the other doses (i.e., D_c, D_d, D_e and D_r), a slight difference between the classes can be noted (as reported in the cases of the energy dissipation dose and radiated energy dose in figures 3.51 and 3.52 respectively); with a minimum time interval of 95 minutes, the energy dissipation dose ranges from 18.379 to 331.77 J/m^3; while, the radiated energy dose varies from 0.12 to 18.8 mJ. Although, the subjects within a certain class do not seem to show a common trend, a gap between the two bands (HS and MS) is observed. The blue band corresponding to the confidence interval of HS is above the red band of MS patients; this means, that healthy subjects demonstrate both, an amount of heat produced in the vocal folds during vibration (D_e) and an energy radiated out of mouth (D_r), that is slightly higher in respect to patients.

In addition, it is possible to notice that subjects often exceed the confidence interval (set to $\pm 1^*\sigma$); moreover, to conduct this investigation subjects (both MS and HS) displaying abnormal doses values are removed (since they are considered outliers) and, among the ones remained, the mean value and standard deviation

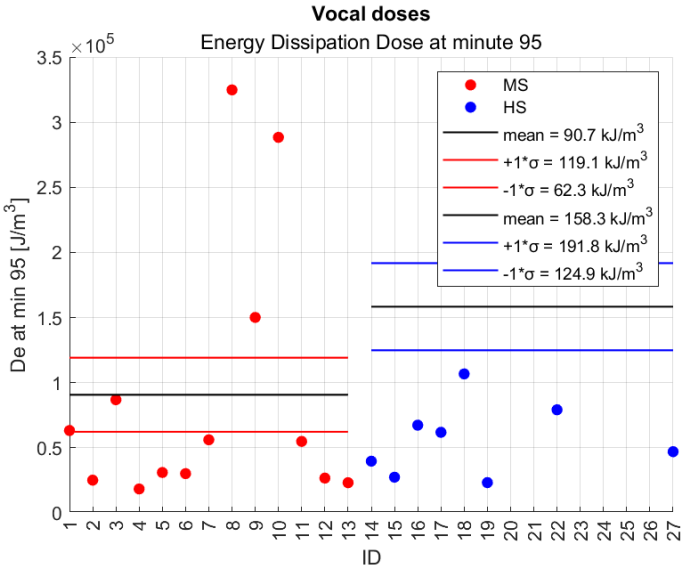

Figure 3.51: Energy dissipation dose values at the minimum time interval of 95 minutes for each subject under analysis

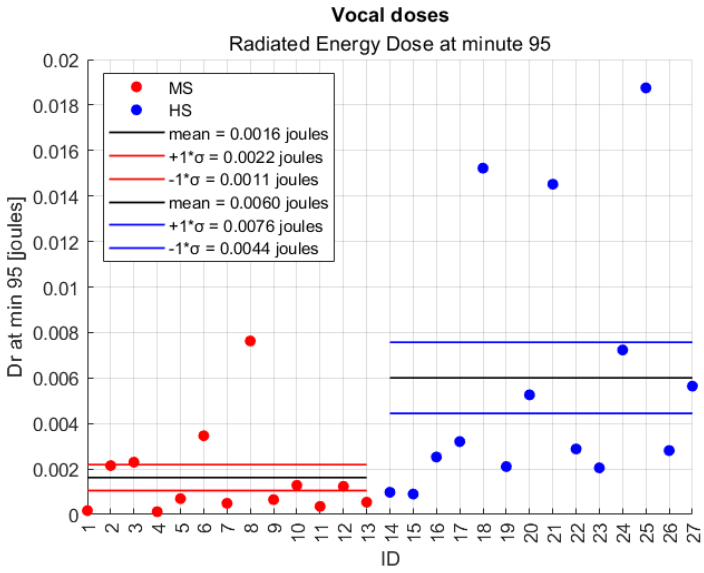

Figure 3.52: Radiated energy dose values at the minimum time interval of 95 minutes for each subject under analysis

are calculated. Also, in this analysis, the two main limitations for the long-term recordings are found, which are the limited data-set (i.e., low number of subjects)

and the time interval too short to assess fatigue.
Furthermore, another analysis performed, consist in weighting the other doses in respect to the time dose, used as a normalization factor to obtain doses per second of vocalization; however, being all the vocal doses, extracted for a time interval of 95 minutes, then, the result is the same if the doses are referred to the time dose instead of the Voicing Time Percentage (PPT) (since the considered interval of time is the same). The idea, in this investigation, is to find a correlation between time dose and the other doses. If no correlation is observed, such as in the case of D_e as a function of D_t (figure 3.53), it means that the value of the energy dissipation dose, among different subjects, at minute 95 is similar. If no significant dependence with respect to the time dose is demonstrated, then, data can be observed without performing normalization (with respect to D_t). A strong correlation between the

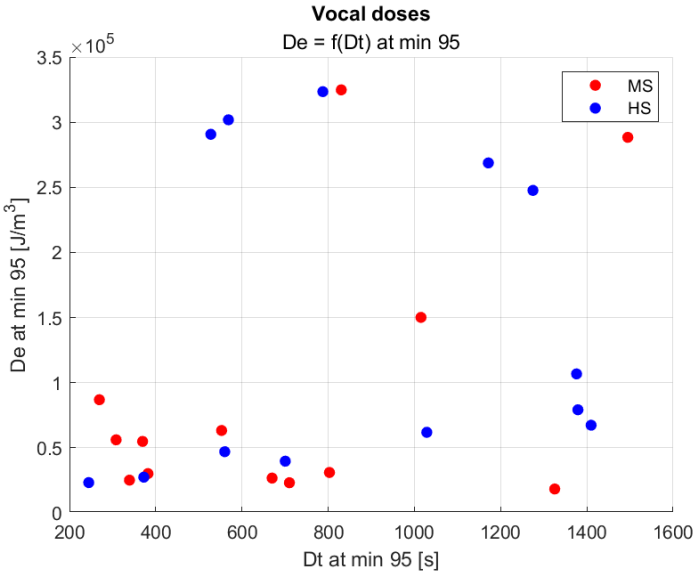

Figure 3.53: Energy dissipation dose values weighted in respect to the time dose at the minimum time interval of 95 minutes

cycle dose and the time dose is expected, since the only variability is introduced by f_0 parameter (as reported in their definition in section 2.2.4). Similarly, a correlation between the distance dose and the time dose is also found (in this case, slightly less in respect to D_c), since variability is introduced from both the parameters f_0 and SPL (this latter defined in the empirical rules of the parameters A, P_{th} and P_L).

It is clear, that, when observing the cycle and the distance dose (figures 3.54 and 3.55), the parameter D_t enters directly into their definition and correlation is found; while for other parameters, such as the energy dissipation dose (figure 3.53)

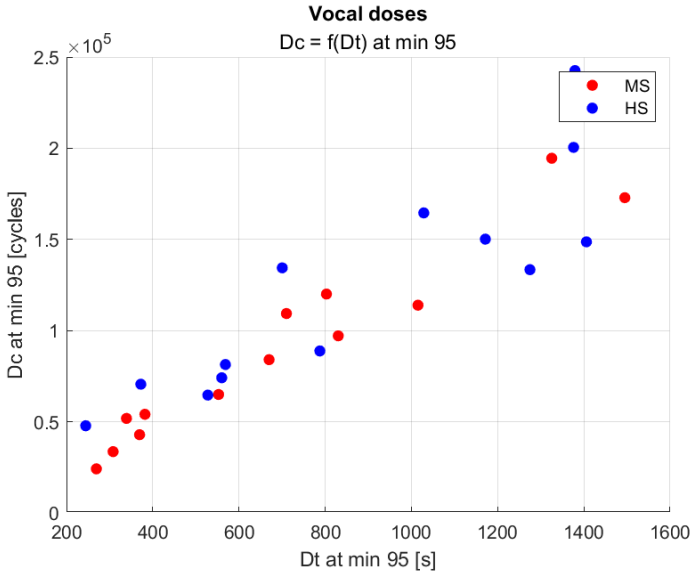

Figure 3.54: Cycle dose values weighted in respect to the time dose at the minimum time interval of 95 minutes

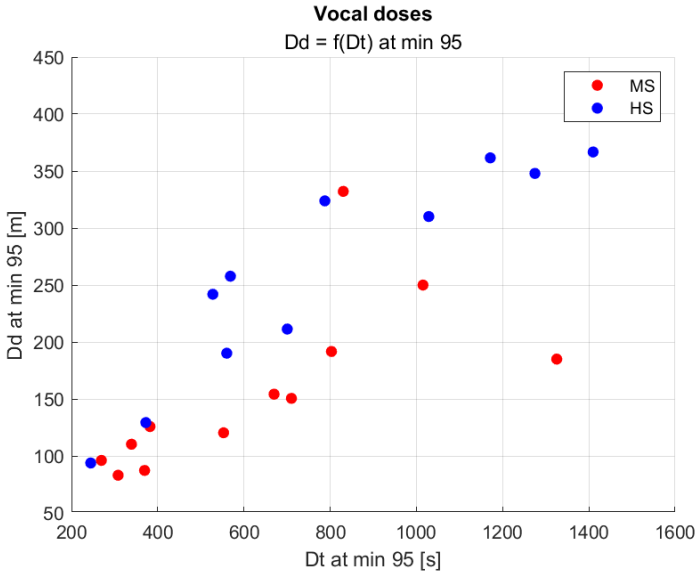

Figure 3.55: Distance dose values weighted in respect to the time dose at the minimum time interval of 95 minutes

or the radiated energy dose, no correction (normalization) is needed.
Since the limit concerning time interval duration too short for long-term assessment

is found, the same study is conducted by increasing the time interval and by eliminating the subjects with a short duration. Considering a minimum time interval of 95 minutes, the total number of subjects in the study are twenty-seven (1-13 MS and 14-27 HS). The data-set comprises the long-term acquisitions of thirty subjects; however, three subjects are considered outliers and therefore, removed. The same analysis conducted with a minimum time interval of 95 minutes, is performed, choosing a minimum time interval equal to 156 and 200 minutes. This choice is motivated by the fact that, the effect of fatigue experienced by the subjects, is most visible as the time interval duration for the long-term monitoring increases. As it is expected, by considering a minimum time interval of 156 min, the classes show a different behavior in the values of vocal doses; observing the results (in figure 3.56), it is clear how much the confidence interval associated with HS is significantly higher in respect to MS patients, as reported in the case of the radiated energy dose (at minute 156). As already mentioned, this dose is related with the efficiency in voice production, which seems much higher in the healthy group. Similarly, weighing D_r as a function of the time dose, an higher trend in the case of HS group in respect to the red elements is shown (as represented in figure 3.57).

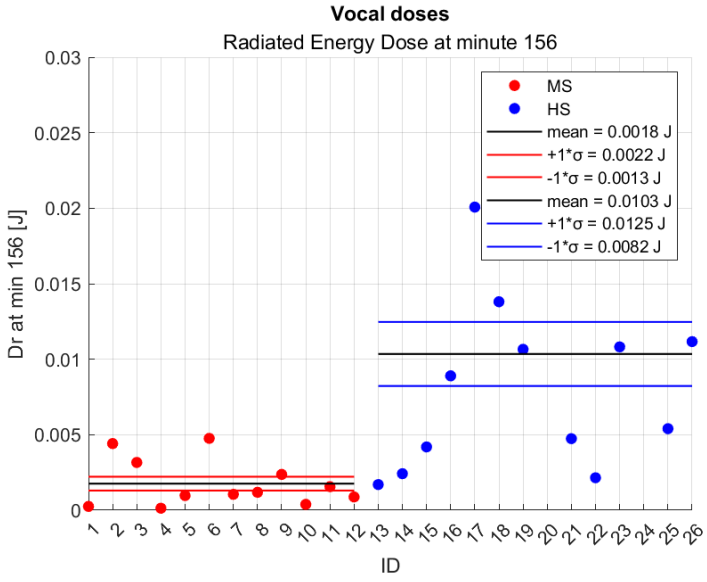

Figure 3.56: Radiated energy dose values at the minimum time interval of 156 minutes for each subject under analysis

Then, increasing the minimum time interval up to 200 minutes, different behavior between classes is observed, as reported for the example of the distance dose (figure 3.58). The total number of subjects in the 200' evaluation are 25 (1-11 MS and 12-25 HS). Eventually, although these measures (i.e., the distance dose

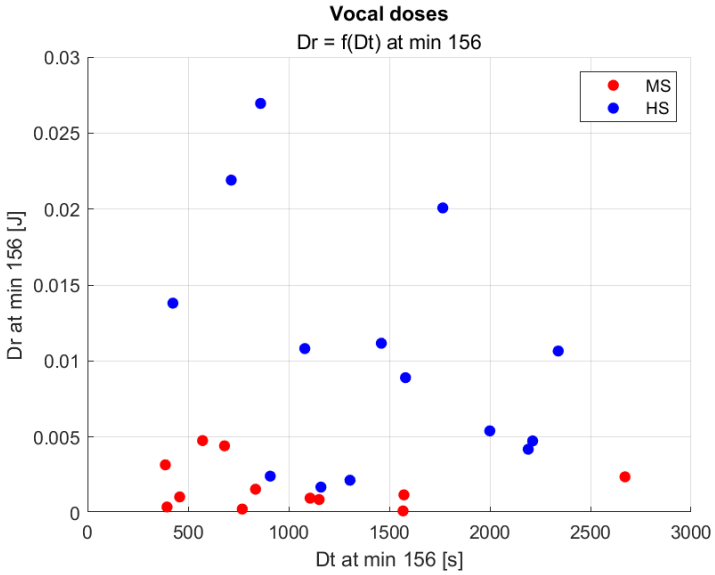

Figure 3.57: Radiated energy dose values weighted in respect to the time dose at the minimum time interval of 156 minutes

and the time dose) are correlated, MS patients show a value of total distance accumulated by the vocal cords during vibration, always lower if compared to the healthy subjects (figure 3.59).

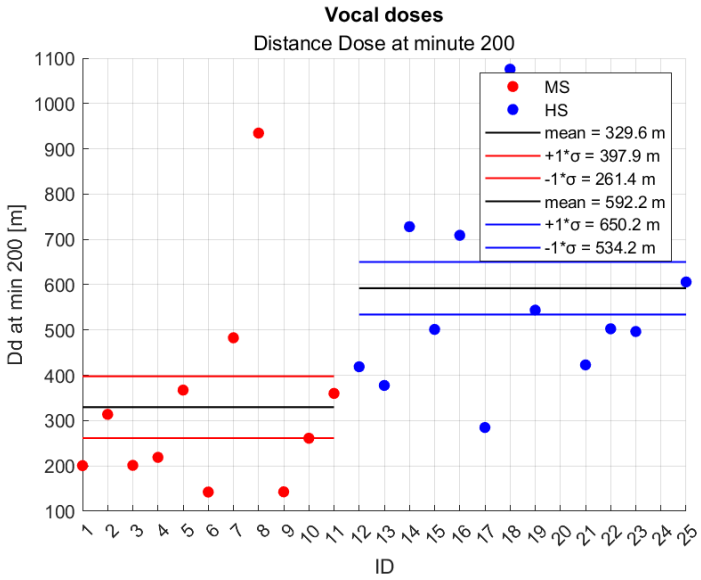

Figure 3.58: Distance dose values at the minimum time interval of 200 minutes for each subject under analysis

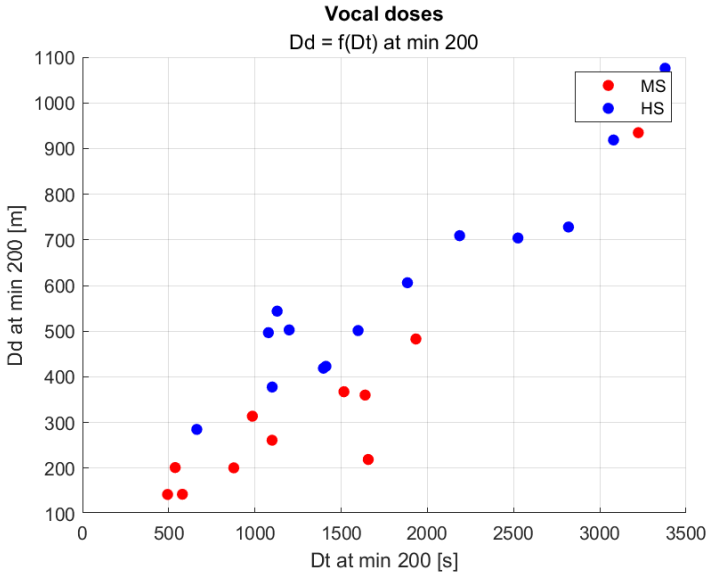

Figure 3.59: Distance dose values weighted in respect to the time dose at the minimum time interval of 200 minutes

Chapter 4

Conclusions

In this work, different studies are conducted on voice recordings of two balanced subgroups of identical dimension (i.e., sixteen subjects), correspondent to healthy subjects (HS) and patients with Multiple Sclerosis (MS). For each subject, speech material consists in three vocal tasks (vocalization of the sustained vowel /a/, reading of a phonetically balanced text and performing approximately one minute of free speech), simultaneously acquired using an in-air microphone system and a contact microphone-based device (Vocal Holter, VH). In addition, long-term recordings are carried out with VH device only, covering a maximum period of four hours of subjects' daily activities. The available vocal material recorded with the microphone in air is pre-processed according to the Harmonic-to-Noise Ratio (HNR) method, in order to select the harmonic frames used to extract vocal parameters in the time, spectral and cepstral domains. The subjects are classified comparing the probability, returned by a Logistic Regression (LR) model, to a fixed threshold (set to 0.5) and dividing them into the classes of HS and MS. The LR model is trained using a single and a combination of 2, 3, 4 features, and the combinations of features (or single feature) providing the best classification performance (in terms of accuracy) are selected. Then, the selected features are used to validate the LR model with 5-fold cross-validation, in the Classification Learner App in Matlab (R2022b). If there are multiple features with the same maximum accuracy value, validation is performed with both the features selected by the algorithm and the features with the largest Area Under The Curve (AUC), among those with the highest accuracy. The best classification results achieved, demonstrate that the balanced speech task is the most suitable vocal material for discriminating between healthy and pathological voices; by selecting 3 features, which are gender, 5° percentile of Cepstral Peak Prominence Smoothed, and harmonic frames ratio V/uV, an accuracy value equal to 92.3% is reached. An unusual result that is found in this investigation, is the presence of the *gender* parameter, considered (among the other parameters) a significant feature for classification. As a result, a

test is conducted by selecting and evaluating the classification performance for the feature combination $f_{0,\text{mean}}$ (instead of gender), $\text{CPPS}_{5,\text{prc}}$ and V/uV, since an high correlation between these parameters is expected; the results obtained, are lowered (showing an accuracy equal to 88.5%), but still express how much this combination of features is significant in distinguishing between the classes. In the other cases of sustained vowel /a/ and free speech task the classification performance are lower in respect to the reading task; with more detail, the selected features in these tasks are: for sustained vowel the $\text{CPPS}_{\text{skewness}}$ parameter alone (accuracy equal to 77.8%) and in the case of free speech task the feature combination consists of HNR_{std}, $\text{CPPS}_{5,\text{prc}}$ and V/uV (accuracy equal to 89.3%).

The expanded uncertainty $U(p)$ of the probability p for each task is evaluated, thus providing a confidence interval, which is created by applying a coverage factor of 2 to the standard uncertainty. When the confidence interval exhibited values that intersect the discrimination probability set to 0.5, the classification of the subject is considered too doubtful, hence the third class of "non-classified" is introduced. In order to get objective feedback on the effect of "non-classified" subjects on overall classification performance, new classification metrics, such as the Realistic Accuracy and the Fraction of Classified (FoC), are defined. The implementation of this procedure to the feature combination showing the best performance during cross-validation (gender, $\text{CPPS}_{5,\text{prc}}$ and V/uV), results in FoC of 92.3% and an higher accuracy.
With the aim of validating VH device, considered a valuable aid for long-term monitoring of fatigue, the parameters extracted from the microphone in air are compared to the ones from VH by calculating differences Δ between these measures. Considering sustained vowel /a/ task, the analysis is performed on the parameters local jitter (%), local shimmer (%), $\text{CPPS}_{\text{median}}$ (dB) and CPPS_{std} (dB); for balanced and free speech task these differences are carried out for all descriptive statistics of fundamental frequency f_0 (Hz) and CPPS (dB). For the parameters local jitter, $\text{CPPS}_{\text{median}}$ and CPPS_{std} the validation can be considered passed, while for the others, such as local shimmer, significant differences in terms of Δ values are noted. Since the two devices have different characteristics, i.e. in terms of bandwidth, and they receive as input different signals (the vibration induced by vocal folds for the VH device and the air-pressure signal for the in-air microphone), the use of the VH device requires the definition of specific cut-off values for the extracted parameters. Additionally, the VH device is preferable both for convenience in conducting acquisitions and for its insensitivity to other possible sound sources. In addition, an attempt to improve the results of Δ is applied by removing subjects showing abnormal behavior (i.e., outliers), and among these remained, a new mean value is calculated; however, also in this analysis, the differences performed between the values of delta shimmer, returned by the two devices, remain high.
A proposal to assess fatigue is conducted with the use of differences δ between the

parameters extracted from the long-term and the short-term monitoring, the last considered as a sort of "baseline"; this comparison is carried out considering the parameters Sound Pressure Level (dB), fundamental frequency f_0 (Hz), CPPS (dB), and Background Noise Level (90° percentile) in dBA. No significant difference in the behavior of the classes with regard to the fatigue experienced is found; however, the main limitations derive from the number of subjects involved that is low, and in the time interval of the acquisitions, too short (between 95 and 200 minutes) to compute an assessment on vocal fatigue. It is expected that, as the time proceeds, a distinction between healthy subjects and patients will emerge, being fatigue one of the most obvious and debilitating symptoms of Multiple Sclerosis. Although the values of δ do not exhibit important differences among the two recordings, the results are consistent, since an increase in fundamental frequency and sound pressure level after long periods of voice usage is found. Additionally, the parameters acquired with VH for the long-term evaluation are visualized over time and for each subject an analysis on the changes in vocal production caused by environmental factors is performed. The correlations between the parameters SPL_{mean}-BNL_{LAF90}, $f_{0,\text{mean}}$-BNL_{LAF90} and SPL_{mean}-$f_{0,\text{mean}}$ are observed by performing an inter-class and an intra-class comparison, and for each of these maps, the regression line and the correlation coefficient R^2 are considered. While in the case of subjects taken individually, no significant difference is noticed, considering the entire HS/MS class, the results are in accordance with what expected; in fact, the MS class demonstrate lower ability than healthy subjects to increase their speech intensity level, as the background noise increases. Since an increase in the noise level, can lead the subject to increase the speech intensity of voice and thus, to hide the effect of fatigue; a compensation on the SPL_{mean} value over time with respect to the noise level is operated by means of the angular coefficient of the regression line modelling the correlation between these parameters. A difference in the value of slope of the regression line, that fits SPL_{mean} "corrected" over time, between the two classes is not observed; also, this result can be explained by observing that the MS group exhibits a time interval duration for the long-term evaluation that is lower in respect to the HS case. Furthermore, MS patients show a lower ability to vary both speech intensity level and fundamental frequency than HS; to assess the variability of SPL_{mean} and f_0 for each group, the standard deviation of these parameters is considered, demonstrating in the MS a lower range of variability.
Eventually, an evaluation of five vocal dose measures as indicators of long-term vocal folds tissue exposure to vibration is provided, which are time dose, cycle dose, distance dose, energy dissipation dose and radiated energy dose. Unless vocal effort is significant among the other subjects, there is the problem of comparing the vocal doses. To compute an assessment at consistent times, a minimum time interval duration common to all subject is considered. Considering a minimum time interval of 95 minutes, while the time dose is not significant in distinguishing the two classes,

in the case of the other doses a slight difference is noted. All the other doses are weighted in respect to the time dose, used as a normalization factor in order to find a correlation. If no correlation is observed, such as in the case of the energy dissipation dose as a function of the time dose, it means that the value of that dose among the subjects at minute 95 is similar and data can be observed without performing a normalization (with respect to the time dose). However, a strong correlation between the cycle and the time dose and also, the distance and the time dose, is found, since the time dose enters directly in the definition of these doses (i.e., cycle and distance dose). Since the limit concerning time interval duration too short for long-term assessment is found, this study is conducted by increasing the time interval and by eliminating the subjects with a short duration. Considering a minimum time interval duration equal to 156 minutes and 200 minutes, different behavior between the classes is shown for all the doses. In the case of the radiated energy dose, the healthy group exhibits significantly higher values in respect to MS, demonstrating that the efficiency in voice production is lower in patients. Also in this investigation, limitations can be overcome through both an increase in the data-set and in the time interval of the records, being the considered acquisitions too short to demonstrate fatigue.
To clarify, this thesis work is performed with the objective of validating the use of VH as a tool, which on the same level as the in-air microphone, is useful in providing distributional parameters able to characterize vocal health and, at the same time, in computing an assessment for long-term monitoring of fatigue.

Appendix A

Appendix

A.1 Anatomy of the Phonatory system

Understanding voice production and voice control, requires an integrated approach, in which physiology, vocal fold vibration, and acoustics are considered as a whole, instead of disconnected components. There are three systems acting together to produce voice, that are, the Respiratory system, the Phonatory system (also know, as "voice box") and the Resonatory system [30]. The lungs are the main organs of the Respiratory system and they are considered as the "power" behind voice production, since they cause the airflow to pass through the vocal folds; as a consequence, their vibration occur and the sound source is created. Vocal folds (or vocal cords) are layered muscles (about 11–15 mm long in adult women and 17–21 mm in men) located in the *larynx* (i.e., the Phonatory system) at the top of the *trachea*. The vocal folds stretches across the larynx along the anterior-posterior direction, attaching anteriorly to the *thyroid cartilage* and posteriorly to the anterolateral surface of the *arytenoid cartilages* (figure A.1). The vocal cords are part of the *glottis*, which are a portion of the laryngeal cavity formed by the vocal folds and the *rima glottidis* (i.e., an opening between them). The phonation cycle is divided into an opening phase (in which the glottis opens), determining the separation of the vocal folds, and the closing phase (i.e., the glottis closes), in which the space between the vocal folds reduces. When the airflow (produced by the lungs) creates pressure below the glottis, the vocal folds are in complete adduction and it coincides with the closing phase. The vocal cords remain closed until a negative intra-glottical pressure is produced, then they are pulled back and thus, opened. The Resonatory system includes the vocal tract from the trachea to the mouth and it is responsible in shaping the tone of the voice; in addition, all the organs in the oral and the nasal cavity play an active role in the generation of consonants and in the production of human voice.

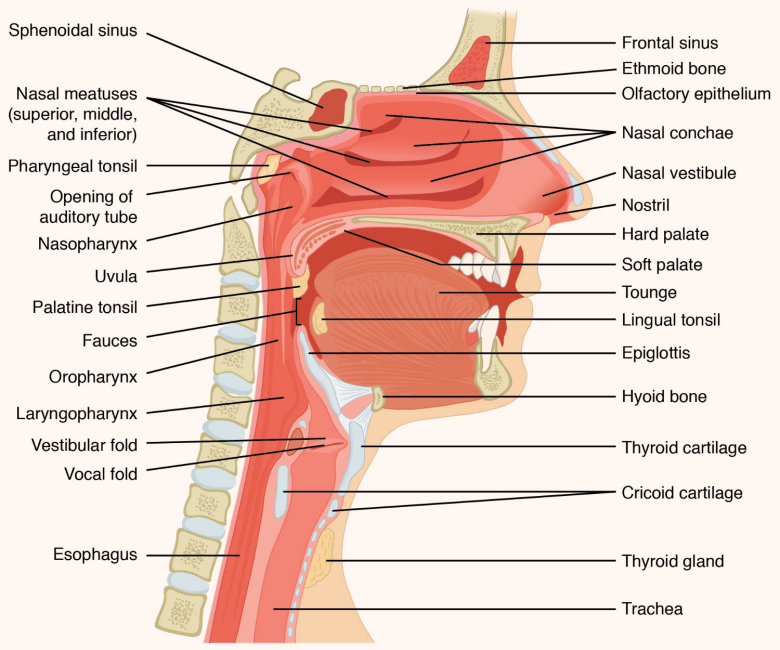

Figure A.1: Anatomical description of the larynx

The speech signal is a complex signal, that has two main components: one related to the glottal pulses and another related to the vocal tract. In vocal production, sounds can be distinguished in voiced and unvoiced sounds. For voiced sound production, vocal fold vibration modulates airflow through the glottis and produces sound, which propagates through the vocal tract; these sounds are characterized by the fundamental frequency, determined by the opening and closing of the vocal folds, and other frequencies (i.e. formants) generated by resonant cavities. Typical examples of voiced sounds are originated by vowels, such as /a/ and /i/. In contrast, unvoiced (or voiceless) sounds, are produced without vocal fold vibration (e.g., sounds generated by airflow through constrictions in the vocal tract or other sound producing mechanisms such as whispering) [31]. In Italian language, unvoiced sounds can be produced by the consonants /k/, /f/ or /t/.

A.2 "Notturno"

Notturno. Vi è un profondo silenzio nel buio della notte. Vincino al pozzo, nella cui acqua si specchiano la luna ed una scia di stelle, la magnolia stende i suoi rami, cespugli di rose olezzano nell'aria. Il temporale è cessato e la pioggia, ormai, non cade più. Solo le rane gracidano nei fossi oltre quel prato.

A.3 Logistic Regression

Logistic regression (LR) is a non-linear statistical model, belonging to the class of Generalized Linear Models, used to separate binary variables as in the case of a pathological group in respect to a healthy control one. Unlike linear regression model, which returns values belonging to the set of real numbers, LR model provides a probability range bounded between 0 and 1. The LR model implements the logistic function to model a binary dependent variable, this function transforms a linear combination into the desired target function and for this reason is called *link function*. The logarithm of the odds (log-odds) for the positive class is a linear combination of independent variable X_i called predictors and regression coefficients β_i, where β_0 is the intercept and $i = 1,2,...N$ with N the number of observations.

$$\begin{aligned} \log(\frac{p}{1-p}) &= \Theta^T \cdot x \\ \Theta^T \cdot x &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_N x_N \end{aligned} \tag{A.1}$$

The probability p returned by the model is a continuous function that can be described with the sigmoid function in Eq. (A.2), which is obtained by inverting the Eq. (A.1).

$$p = \frac{\exp^{(\Theta^T \cdot x)}}{1 + \exp^{(\Theta^T \cdot x)}} = \frac{1}{1 + \exp^{-(\Theta^T \cdot x)}} \tag{A.2}$$

The aim of the LR in binary classification problems, is to reduce the distance between the sigmoid function, defined by the regression coefficients $\beta_0, \beta_1, ...\beta_N$ and the ideal step function between 0 and 1. During the phase of the model training, the best coefficients or weights are calculated: this can be done by solving a minimization problem on the log-likelihood through the use of deterministic or stochastic mathematical approaches, such as the least square differences, the gradient descent and the Newton method [14]. The learning algorithm performs the search of the best regressed coefficients and also, gives an estimate of the coefficients' variances and covariances which can be used in order to evaluate the goodness of the regression model. The model provides a continuous probability p of belonging to the positive class. To obtain a binary classification, the probability is compared to a fixed threshold typically equal to 0.5. LR is a powerful supervised Machine Learnine (ML) algorithm, very popular for binary classification problems. The algorithm is trained in order to produce predictions with a dataset of known features and responses. The probabilities returned by the model are compared with the known responses (the true class) to obtain the quantities described by a *confusion matrix*. A confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of a classifier when supervised training is applied. Confusion matrices may be used with any number of classes, but when the classifier deals with only two classes

(positive and negative), the following quantities are derived. True Negatives (TN) is the number of negative elements correctly classified. True Positives (TP) are the number of positive elements correctly classified. False negatives (FN) are the number of positive elements misclassified (i.e., assigned to the negative class). False Positives (FP) are the number of negative elements misclassified (i.e., assigned to the positive class). These quantities are useful to evaluate metrics to assess the performance of the classifier. The following metrics are the most common:

- Accuracy:
$$\frac{TP + TN}{TP + TN + FP + FN} \tag{A.3}$$
it is the most important parameter in classification models and it provide an assessment of the proportion of elements correctly classified.

- Precision:
$$\frac{TP}{TP + FP} \tag{A.4}$$
it represents the probability of an element being positive, when it is classified as positive and it indicates the model's ability to classify positive elements.

- Sensitivity:
$$\frac{TP}{TP + FN} \tag{A.5}$$
also known as true positive rate (TPR), it measures the fraction of positive elements correctly classified. According to the aim of this work, a high TPR corresponds to a low number of FN (i.e., pathological subjects classified as healthy), which means that the classification model has a good ability to identify patients.

- Specificity:
$$\frac{TN}{TN + FP} \tag{A.6}$$
also called true negative rate (TNR), it expresses the proportion of negative elements correctly classified. For a classifier, a high specificity translates into a good ability to recognise healthy subjects. As a result, another metric can be introduced, which is False Alarm = 1 – Specificity.

- Area Under the ROC Curve (AUC): it is another effective measure of a classifier accuracy. The receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. The ROC curve shows the Sensitivity (TPR) against the False Alarm (FPR) at various threshold settings (figure A.2).

AUC can be interpreted as the average value of sensitivity for all the possible values of specificity and it is estimated in Matlab environment throughout the built-in function *perfcurve*. AUC values vary from 0, which means test incorrectly classify all elements, to 1, which is the maximum and it means that the diagnostic test is perfect in the differentiation between the two classes; when the AUC is 0.5, means random discrimination.

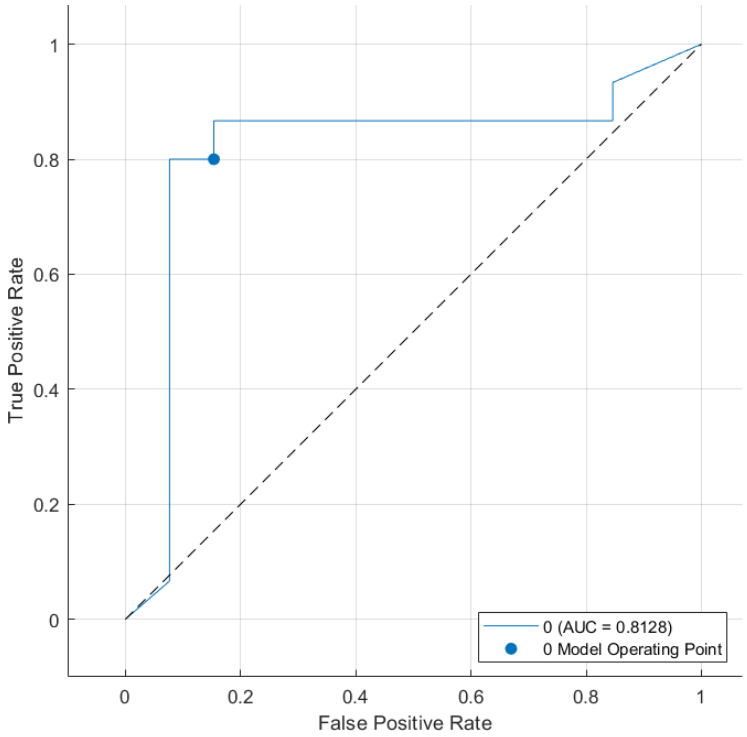

Figure A.2: Example of ROC curve computed by the Classification Learner App in Matlab (R2022b)

Bibliography

[1] «Treating speech problems». In: (2023). URL: https://www.nationalmssociety.org/Symptoms-Diagnosis/MS-Symptoms/Speech-Disorders#section-2 (cit. on p. 2).

[2] Bassem Yamout, Nabil Fuleihan, Taghrid Hajj, Abla Sibai, Omar Sabra, Hani Rifai, and Abdul-Latif Hamdan. «Vocal symptoms and acoustic changes in relation to the expanded disability status scale, duration and stage of disease in patients with multiple sclerosis». In: *European Archives of Oto-Rhino-Laryngology* 266 (2009), pp. 1759–1765 (cit. on pp. 2, 27).

[3] «Phonetically Balanced». In: Farlex Partner Medical Dictionary (Aug. 2012). URL: https://medical-dictionary.thefreedictionary.com/phonetic+balance (cit. on p. 5).

[4] «Software Instruction Manual Multi-Dimensional Voice Program (MDVP) Model 5105». In: (Sept. 2005), pp. 135–189 (cit. on p. 8).

[5] Paul Boersma et al. «Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound». In: *Proceedings of the institute of phonetic sciences*. Vol. 17. 1193. Amsterdam. 1993, pp. 97–110 (cit. on p. 9).

[6] David A Puts, Coren L Apicella, and Rodrigo A Cárdenas. «Masculine voices signal men's threat potential in forager and industrial societies». In: *Proceedings of the Royal Society B: Biological Sciences* 279.1728 (2012), pp. 601–609 (cit. on pp. 9, 45).

[7] Farzad Izadi, Ramin Mohseni, Ahmad Daneshi, and Nazila Sandughdar. «Determination of fundamental frequency and voice intensity in Iranian men and women aged between 18 and 45 years». In: *Journal of Voice* 26.3 (2012), pp. 336–340 (cit. on p. 9).

[8] Alan V. Oppenheim and Ronald W. Schafer. «From frequency to quefrency: A history of the cepstrum». In: *IEEE signal processing Magazine* 21.5 (2004), pp. 95–106 (cit. on p. 10).

[9] Yolanda D Heman-Ackah, Deirdre D Michael, and George S Goding Jr. «The relationship between cepstral peak prominence and selected parameters of dysphonia». In: *Journal of Voice* 16.1 (2002), pp. 20–27 (cit. on p. 10).

[10] Antonella Castellana, Alessio Carullo, Simone Corbellini, and Arianna Astolfi. «Discriminating pathological voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel». In: *IEEE Transactions on Instrumentation and Measurement* 67.3 (2018), pp. 646–654 (cit. on pp. 11, 61).

[11] Rubén Fraile and Juan Ignacio Godino-Llorente. «Cepstral peak prominence: A comprehensive analysis». In: *Biomedical Signal Processing and Control* 14 (2014), pp. 42–54 (cit. on p. 11).

[12] Antonella Castellana et al. «Towards vocal-behaviour and vocal-health assessment using distributions of acoustic parameters». In: (2018) (cit. on pp. 12, 27, 30, 35).

[13] E Chandra Blessie and E Karthikeyan. «Sigmis: A feature selection algorithm using correlation based method». In: *Journal of Algorithms & Computational Technology* 6.3 (2012), pp. 385–394 (cit. on p. 16).

[14] Giulia Resio. «Extraction and selection of vocal features for the assessment of surgeries and rehabilitation of post laryngectomy patients». PhD thesis. Politecnico di Torino, 2022 (cit. on pp. 17, 18, 48, 92).

[15] Alessio Carullo, Alberto Vallan, Marco Fantini, and Giovanni Succo. «Vocal-Feature Based Classification of Post-Laryngectomy Patients for Rehabilitation Monitoring». In: *IEEE Transactions on Instrumentation and Measurement* (2023) (cit. on p. 18).

[16] Alessio Atzori et al. «Effects of measurements uncertainty on classification algorithms, as applied to vocal features for health assessment and early diagnosis». In: (2022) (cit. on p. 18).

[17] «Vocal Holter Med Instruction Manual». Version 2.1.0 version. In: (2018), pp. 1–24 (cit. on p. 20).

[18] A Carullo, A Vallan, A Astolfi, L Pavese, and GE Puglisi. «Validation of calibration procedures and uncertainty estimation of contact-microphone based vocal analyzers». In: *Measurement* 74 (2015), pp. 130–142 (cit. on p. 20).

[19] Birgitta Berglund, Thomas Lindvall, Dietrich H Schwela, World Health Organization, et al. «Guidelines for community noise». In: (1999) (cit. on p. 25).

[20] Pasquale Bottalico, Rachael N Piper, and Brianna Legner. «Lombard effect, intelligibility, ambient noise, and willingness to spend time and money in a restaurant amongst older adults». In: *Scientific reports* 12.1 (2022), p. 6549 (cit. on p. 26).

[21] Marisa P McGinley, Carolyn H Goldschmidt, and Alexander D Rae-Grant. «Diagnosis and treatment of multiple sclerosis: a review». In: *Jama* 325.8 (2021), pp. 765–779 (cit. on p. 27).

[22] Alessio Carullo, Alberto Vallan, and Arianna Astolfi. «Design issues for a portable vocal analyzer». In: *IEEE Transactions on instrumentation and measurement* 62.5 (2013), pp. 1084–1093 (cit. on p. 36).

[23] Ingo R Titze, Jan G Svec, and Peter S Popolo. «Vocal dose measures». In: (2003) (cit. on pp. 36, 75).

[24] Jan G. Švec, Peter S. Popolo, and Ingo R. Titze. «Measurement of vocal doses in speech: experimental procedure and signal processing». In: *Logopedics Phoniatrics Vocology* 28.4 (2003), pp. 181–192 (cit. on pp. 38, 75).

[25] ANSI S3. 5-1997. «S3.5-1997, Methods for the calculation of the speech intelligibility index». In: *New York: Amer. Nat. Standards Inst* 19 (1997), pp. 90–119 (cit. on p. 38).

[26] Jan G Švec, Ingo R Titze, and Peter S Popolo. «Estimation of sound pressure levels of voiced speech from skin vibration of the neck». In: *The Journal of the Acoustical Society of America* 117.3 (2005), pp. 1386–1394 (cit. on p. 38).

[27] Sara Palmieri. «Assessment of the vocal status of multiple sclerosis patients-comparison with healthy subjects and evaluation of vocal rehabilitation». PhD thesis. Politecnico di Torino, 2023 (cit. on p. 48).

[28] «10^{th} Convention of European Acoustic Association». In: Turin, Italy, 2023 (cit. on p. 59).

[29] Andrea Nacci, Bruno Fattori, Valentina Mancini, Erica Panicucci, Francesco Ursino, FM Cartaino, and S Berrettini. «The use and role of the Ambulatory Phonation Monitor (APM) in voice assessment». In: *ACTA otorhinolaryngologica italica* 33.1 (2013), p. 49 (cit. on p. 64).

[30] «How Does My Voice Work?» In: (Apr. 2018). URL: `https://www.templehealth.org/about/blog/how-does-my-voice-work` (cit. on p. 90).

[31] Zhaoyan Zhang. «Mechanics of human voice production and control». In: *The journal of the acoustical society of america* 140.4 (2016), pp. 2614–2635 (cit. on p. 91).