POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering
Biomedical Instrumentation

Master's Degree Thesis

Classification of multiple sclerosis patients through vocal features

Performance evaluation by using different vocal indexes and software applications

Supervisors	Candidate
Prof. Alessio CARULLO	Federica SECUNDO
Prof. Alberto VALLAN	

DECEMBER 2023

Summary

Multiple sclerosis (MS) is a chronic disease of the central nervous system that affects the brain and spinal cord. MS is an autoimmune disease, meaning that the body's immune system mistakenly attacks its own healthy tissue, which in the case of MS is myelin, a substance that lines nerve fibres and helps transmit nerve impulses efficiently. MS involves various body activities of patients, including language. This study was conducted in collaboration with Don Gnocchi Foundation in Milan. Vocal recordings of 16 subjects with MS and 16 subjects (HS) without MS were acquired, which include 3 repetitions of the vowel /a/, free-speech of about 1 min, and the reading of a phonetically balanced text. For each subject, an air microphone (MI) and a contact microphone (VH) were used to simultaneously acquire the vocal signals. Only for the MI, the available traces were manually analysed to exclude invalid recordings (saturated or too noisy) and select the parts of interest. Then, Matlab scripts were specifically developed to subdivide the vocal signal in frames and extract the parameters Harmonic to Noise Ratio (HNR), Cepstral Peak Prominence Smoothed (CPPS), fundamental frequency (fo), and signal intensity (RMS). Each parameter is represented by means of 9 descriptive statistics (mean, median, mode, standard deviation, range, 5 percentile, 95 percentile, skewness, kurtosis). For the vowel /a/, other 9 stability parameters of amplitude (shimmer) and period (jitter) were extracted. The purpose of this thesis is to identify the parameters that better distinguish MS vs HS classes and two indexes that group some of the extracted parameters have been investigated: the well-known Acoustic Voice Quality Index (AVQI) and the Warning Score (WS), which is a new index proposed to assess the vocal health status of subjects. AVQI depends on the parameters jitter, shimmer, CPPS, HNR, Spectral Slope and Tilt extracted by a concatenation of 3 seconds of sustained vowel /a/ and 3 seconds of

reading. These parameters were extracted through 3 applications (Matlab, Praat and VOXplot) and they show significant differences: in particular, Matlab CPPS mean values differ of about 3 dB compared to Praat. In terms of AVQI, only parameters extracted with Praat and VOXplot were used: subjects were classified using the Logistic Regression (LR) model, by comparing their accuracy (Acc=70.8% with VOXplot and Acc= 61.5% with Praat) and area under curve (AUC=0.63 with VOXplot and AUC=0.54 with Praat). The index WS depends on the parameters local jitter, local shimmer, mean and standard deviation of CPPS extracted using Matlab scripts from the vowel /a/ by both MI and VH. Subject were also classified according to the index WS by the LR (Acc=41.7%, AUC=0.36). Classification results between HS and MS are not outstanding, thus highlighting that even if HS do not have MS, they could be exhibit dysphonic behaviour. For this reason, from a data set of 58 True Healthy Subjects (THS), 12 subjects were extracted in order to have a balanced data set with 12 MS and the classification was repeated using the same LR model, obtaining significantly better results: Acc=100% (AVQI evaluated by VOXplot), Acc=92.3% (AVQI by Praat), Acc=87.5% (WS by Matlab). The LR classification using both AVQI and WS does not provide improvement (Acc=91.7%). Furthermore, a comparison between perceptual assessment (G and A indexes of the GIRBAS scale) and input parameters of the WS index showed negligible correlation both for MS and HS.

Acknowledgements

This study was the result of a collaboration between the Polytechnic of Turin and the team of speech therapists from the Don Gnocchi Foundation Hospital in Milan, who made available the voice recordings of multiple sclerosis patients.

With much appreciation, I would like to thank professor Alessio Carullo, my excellent thesis advisor, for his amazing assistance and support as I worked on this thesis. His knowledge, endurance, and commitment have been crucial in making our job possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Anatomy and Physiology of the Voice Production

The collection of all the structures involved in the production and modulation of phonemes is known as the phonatory apparatus (Figure 1.1) . It is divided into several parts at the level of the mouth, nasal cavities, and neck, and it works during the exhalation phase, requiring very little of the air that is released.
Each component of the apparatus is tailored to perform a certain function, which is invariably related to the creation or modification of sound. The organs of which it consists can be distinguished into:

- The *vocal cords* are located in the larynx, an organ in the front part of the neck.
- The *larynx* is located in the antero-superior part of the neck and in which the vocal cords reside.
- The *nasal cavity*, which is situated above the oral cavity. Along with filtering, warming, and humidifying the air that is inhaled, the nasal cavity also controls how loud sounds vibrate.
- The *oral cavity* contains the tongue, palate, palatine veil, lips, and teeth.
- The *pharynx*, often known as the throat, is a passageway for food and air, contains the tonsils, and serves as a resonance chamber for sounds made.

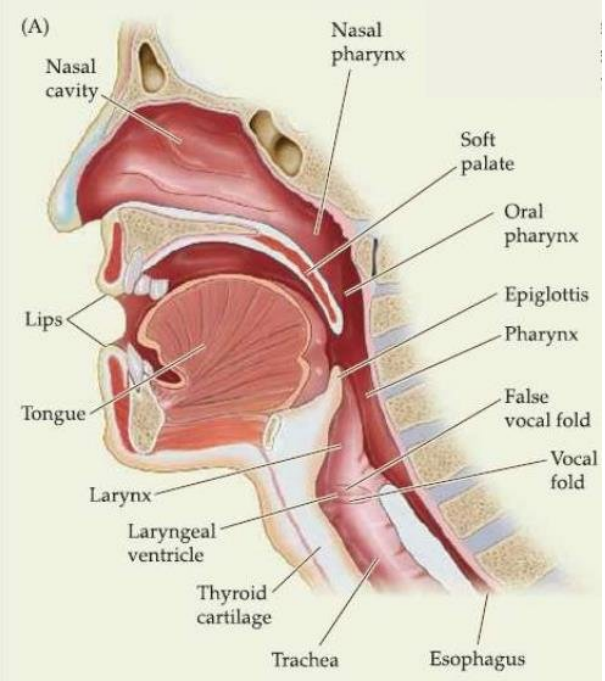

Figure 1.1: Anatomical position of the larynx in the neck.[1]

The larynx is composed of mucosa-coated cartilage and has a triangular pyramidal shape and it is divided in three zones:

1. The *supraglottic zone*, also known as the false vocal cords, is the upper portion between the laryngeal cartilage, also known as the epiglottis, and the upper pair of vestibular folds. The *epiglottis* plays an important role in the protection of the lower respiratory tract: it lets the air in towards the trachea during breathing, but it closes while swallowing, to block food and drinks from going down into the trachea. [2]

2. The *glottis area*, the middle section, is the seat of the true vocal cords, appearing as a pair of flaps formed by a complex of ligaments, muscles, and squamous epithelium that demarcate a variable space, known as the glottis rima (or just glottis). Even though there are four vocal cords, only two are really used for phonation. In fact, two ventricular folds, sometimes known as fake vocal cords, are located over both sides of the glottis and are employed to produce deep sonorous tones. As proof, the ventricular folds appear thinner while the real voice cords appear bigger, surrounded in muscle fibers, and with

a small space between them (Figure 1.2).
The opening phase of phonation, during which the vocal cords are parted (inspiration), and the closing phase, during which the space between the vocal cords is somewhat reduced (exhalation), are the two distinct phases of the phonation cycle. During the closure phase, when the vocal cords are fully adducted, the pressure below the glottis rises as a result of the airflow produced by the lungs. Until the subglottal pressure is high enough to push the vocal cords apart and create a negative intraglottical pressure, which draws the vocal folds back and closes the glottis, the vocal cords remain closed. The cycle is then repeated, resulting in an acoustic wave that travels down the vocal tract between the trachea and the mouth and permits prolonged vibrating of the vocal folds.

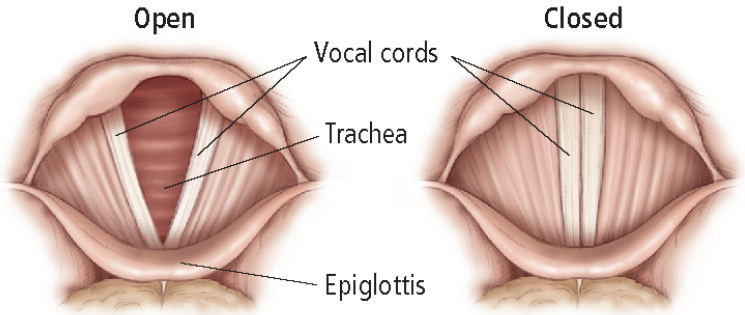

Figure 1.2: View of the interior of the larynx during the opening and closing phase of the vocal folds.

3. The *subglottic area*, or lower segment, connects the trachea directly to the laryngeal cartilage, also known as the cricoid cartilage, from the area immediately after the glottis.

1.2 Voice Signal

The emitted speech signal is a complicated signal that typically consists of turbulent noise produced by airflow passing through various resonant environments and quasi-periodic vibration of the vocal chords. Two distinct sounds can be produced depending on the kind of source that causes the phoneme:

1. Voiced sounds, which are those produced when the vocal cords vibrate as a result of air passing through the glottis, i.e. vowels /a/ and /i/ in the Italian language, originate during exhalation. The fundamental frequency, which is the frequency at which the vocal cords open and close, and the formants, which are the characteristic frequencies around which amplitude peaks in the signal pattern occur and which are brought on by the effects of resonant cavities, are these main characteristics of the objects.

2. Unvoiced sounds, also known as "voiceless consonants", in which turbulence is produced by forcing air through a constriction in the resonant tract rather than using the vocal chords to produce the phoneme [3]. One such is the sound /s/ in the word "silence".

Analyzing a signal in either the time domain or the frequency domain allows for the search for formants and fundamental frequencies. Due to the differing shape of the vocal folds, which are bigger and longer in the case of males, the fundamental frequency F_0 of vocalized sounds oscillates around an average value, which is a characteristic of each individual and varies according to age, gender and type of vocal activity as shown in Table 1.1.

Table 1.1: Frequency range depending on gender [4]

Frequency	*Type*
75 Hz ÷ 300 Hz	Man
100 Hz ÷ 400 Hz	Woman

The tension in the vocal cords, tiny differences in the shape, conformation, or flexibility of the various resonant chambers, or both, are commonly used to produce vocal signals at different frequencies. As a result, it is essential to appropriately analyze the signal by dissecting the different phonatory events and distinguishing between those with high harmonic content and those with more noise, which would skew beneficial information.

1.3 Vocal symptoms and acoustic changes in patients with multiple sclerosis

Speech is a complicated process that involves the coordination of several bodily systems, including the neurologic system, and it is a reflection of how well the body as a whole is doing. The chronic degenerative condition known as multiple sclerosis (MS) damages the myelin sheath, causing many lesions in the brain's white matter, brainstem, and spinal cord that significantly impair one's ability to move. Vocal symptoms and acoustic measures of patients with multiple sclerosis (MS) are investigated in relation to the duration of the disease, stage of the disease and the degree of disability [3].
Dysarthria, a neuromuscular disorder that results in disturbances in the motor control of the speech mechanism and is frequently accompanied by other symptoms brought on by lesions in the brainstem, and dysphonia, a voice disorder, which frequently coexists with dysarthria because the same muscles, structures, and neural pathways are used for both speech and vocal production, are among the symptoms of multiple sclerosis that may manifest. As a result, simultaneous changes in voice quality, nasal resonance, tone control, volume, and emphasis are also possible. There are various types of multiple schelerosis [5]:

- Relapsing Remitting: RR is the type of multiple sclerosis that is most prevalent. This kind, which is marked by acute sickness episodes (also known as "relapses") alternated with periods of full or partial health (also known as "remissions"), affects about 85 % of people who are initially diagnosed. Additionally, the RR form can be classified as active (relapses and/or evidence of disease activity on resonance imaging) or inactive, as well as with worsening (proven increase in impairment for a predetermined amount of time after a relapse) or without worsening.
  Though each person's experience with RRMS will be different, the Figure 1.3 above illustrates the various disease activity that can occur in RRMS. After a relapse, the new symptoms may go away completely without increasing the degree of disability, or they may only go away partially, increasing disability. The arrows indicate new lesions on the MRI, which frequently happen after a relapse. New MRI lesions indicating MS activity, however, can occasionally

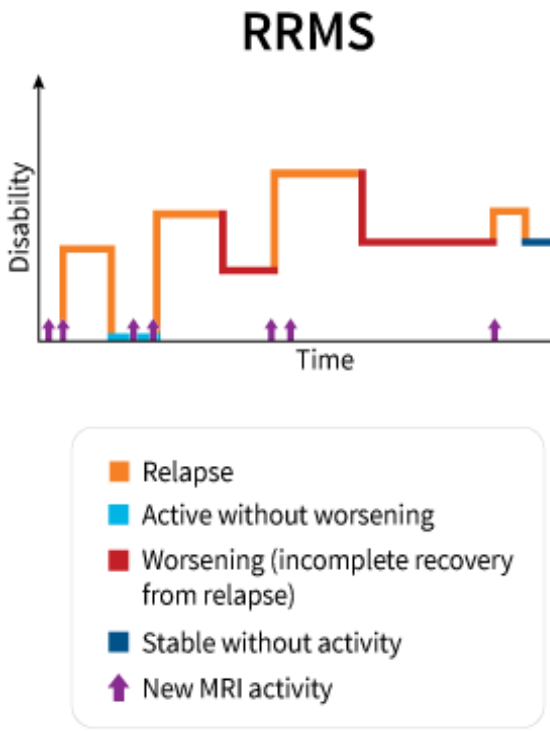

Figure 1.3: RRMS activity [6]

develop without the person exhibiting symptoms.

- Secondarily Progressive: SP is the evolution of the relapsing-remitting form many people who are first diagnosed with the RR form may go on to develop a secondarily progressive form, which is defined by a permanent disability that advances with time. The SP form can also be classified as progressive (objective evidence of disease worsening over time, with or without relapse or signs of disease activity on resonance imaging) or non-progressive. The former refers to the presence of relapse and/or evidence of disease activity on resonance imaging.
  Despite the fact that each person's experience with SPMS is different, the Figure 1.4 illustrates the various disease activities that might manifest in SPMS. The graph shows that after a period of relapsing-remitting illness, the impairment may worsen over time, with or without signs of disease activity (relapses or changes on an MRI). Periods of stability and sporadic relapses

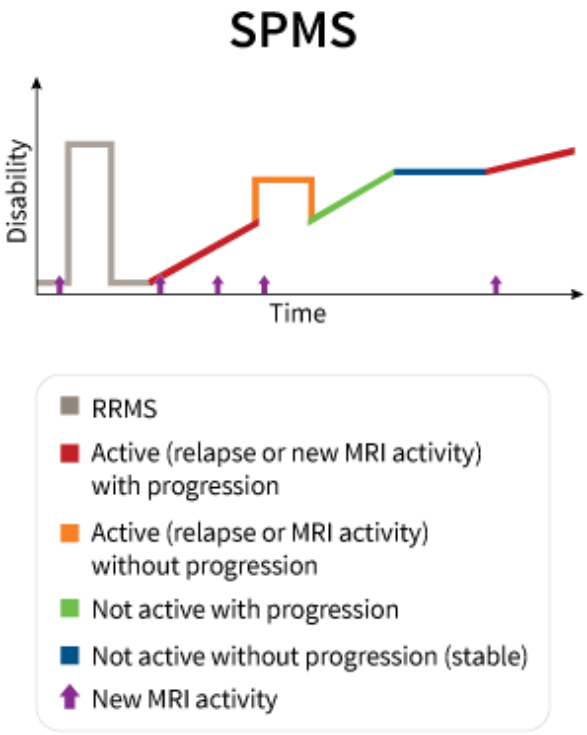

Figure 1.4: SPMS activity [7]

are also possible.

- Primarily Progressive: PP is distinguished by a decline in neurological function that occurs as soon as the initial symptoms appear, without a true relapse or remission. These forms can be categorized as progressive (objective evidence of the disease getting worse over time, with or without relapses or signs of disease activity on resonance imaging) or non-progressive. Active forms are those with occasional relapses and/or evidence of disease activity on resonance imaging. A predominantly progressive variant of multiple sclerosis affects about 15 % of patients.
  The diagram (Figure 1.5) illustrates the numerous disease activities that can occur in PPMS. The diagram shows that there can be short intervals of disease stability, with or without relapses or new MRI activity. Additionally, the patient may go through phases of increasing disability, with or without new relapses or MRI lesions.

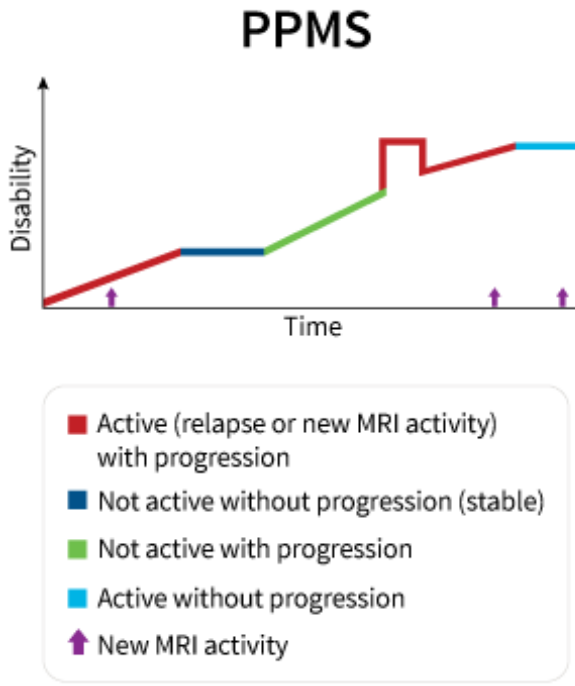

Figure 1.5: PPMS activity [8]

1.4 Perceptual Rating Scales: GIRBAS

Perceptual rating scales are frequently employed in clinical settings to evaluate patient progress made during multiple speech therapy sessions in terms of their phonatory skills. They are composed of a list of vocal characteristics that one or more specialists in the field assign a grade to. A low grade typically indicates a voice or characteristic of high quality, while a high grade typically indicates the low quality. There are scales that patients themselves fill out called quality-of-life questionnaires, from which perceptual information about the subject can be derived. There are scales that can assess phonatory abilities while also determining the status of a certain ailment.
Following are the most typical perceptual rating scales:

- EDSS scale: The Expanded Disability Status Scale is a scale that measures the degree of disability that MS patients have. It has a range of 0 to 10, with 0 representing a normal neurological evaluation. The result is determined by adding the partial results from several functional systems connected to

nervous system activity (such as the pyramidal, cerebellar, sphincter, and other systems). The EDSS is extensively used because it makes it possible to more easily assess how the disease is progressing and to evaluate how well the current course of treatment is working.

- Barthel scale is used to assess the level of independence of the patient. It comprises of ten ADL (Activities of Daily Living) that are common to daily life. The sum of the scores for each item, which might total 100, indicates the patient's level of independence in carrying out activities of daily living.
- Mini-Mental State Examination is a neuropsychological test for determining the existence of cognitive impairment and disorders of intellectual efficiency. The maximum attainable score is 30, if the subject achieves a score below 24 it is defined as pathological.
- GIRBAS scale: Different factors are used to evaluate voice quality as can be seen from the Table 1.2; each factor is given a score between 0 (normal voice) and 3 (pathological voice).

Table 1.2: GIRBAS scale description [9]

COMPONENT	DESCRIPTION
G - Grade	Generic grade of dysphonia;
I - Instability	Voice functionality changes over time, which is crucial for long-term evaluation;
R - Roughness	Low frequency aperiodicity caused by abnormally vibrating vocal folds, which causes variations in wave fundamental frequency and amplitude;
B - Breathy	Unfinished glottis closure, which results in the voice's audible turbulent noise, is created;
A - Asthenic	A feeling of fatigue brought on by insufficient muscle tension caused by a weak voice and a lack of high frequency harmonics;
S - Strained	Evaluation of a hyperfunctional phonetic state characterized by noise, harmonics at high frequencies, and a high fundamental frequency.

Chapter 2

Materials and Methods

This project is done in partnership with the speech therapy and rehabilitation department of Don Gnocchi Hospital in Milan. Don Gnocchi Foundation was born in 1945 thank to Don Carlo Gnocchi and now the foundation operates twenty-five residential institutions and twenty-seven clinics arranged geographically, all of which are supported by the Italian Nation Health Service. The Foundation's goal is to meet the health and care requirements of individuals who are suffering and frail by tending to patients and those who are called to support them, including volunteers, family members, and medical professionals. The Foundation's diverse team of caregivers and medical specialists treats patients of all ages for rehabilitation, finds treatments for children with disabilities of all kinds, looks after elderly individuals who are unable to care for themselves, and attends to patients who are terminally sick. Patients with Multiple Sclerosis (MS), the class of subjects included in this thesis study, are treated by the speech therapy and rehabilitation department of Don Gnocchi Hospital.

Specifically, the effects of multiple sclerosis (MS) on voice quality are examined through the analysis of a data-set containing the voice recordings of thirty-two subjects: sixteen healthy adults (HS) with a mean age of 42 years, standard deviation of 12 years) and sixteen MS patients (mean age 44 years old, with a standard deviation of almost 14 years). The diagram shown in Figure 2.1 provides a basic overview of all the procedures and data addressed in various jobs with the aim of giving a more comprehensive picture of the work.

The main purpose of this study is to compare parameters extracted from the records of patients (MS) and healthy subjects (HS) given by Don Gnocchi's Foundation. For each subject vocal recordings were provided by the in-air microphones (MI) and by contact microphone (VH). MI's vocal recordings were pre-processed to extract parameters. The extracted parameters were evaluated according to two indices: Acoustic Voice Quality Index (AVQI), known in the literature, and Warning Score (WS) which depends on the value of the parameter a score was given. Furthermore, the AVQI index was evaluated with various software applications: Matlab, Praat and VOXplot. Then, this data-set is compared with the true healthy subjects (THS): in fact another data-set available at the Electronics and Telecommunications Department of Polytechnic of Turin have been processed in order to obtain reference data. Finally, all the subjects were classified using the Logistic Regression (LR) model, by comparing some metrics like accuracy and area under the ROC curve.

2.1 Data Acquisition

Voice monitoring has been made possible through the development of technological innovations: the Vocal Holter (VH), portable vocal analyser created at Politecnico di Torino. VH was used to monitor voice quality. With the use of data analysis, it can determine whether a person is vocally healthy, at risk now, or at risk in the future, making it valuable as both a primary prevention electromedical device and a diagnostic tool.

Vocal Holter (VH), in Figure 2.2, is a kit composed by three component: the contact microphone (model hx-505-1-1) that is worn around the neck and detects vibrations caused by the vocal folds during phonation periods; the power adapter and cable; the Data Acquisition and Processing (DAP) unit, which embeds a microphone in air and uses a spacer to keep the subject's mouth at a fixed and known distance from the in-air microphone during calibration. The microphone contains a pin that may slide over another, allowing it to be widened or tightened in accordance with the size of one's neck because every subject is unique from another even physically, as it can be seen in Figure 2.3. When positioning the microphone, care should be made to ensure that it is comfortable so that it will not need to be moved again for the duration of the voice recording.

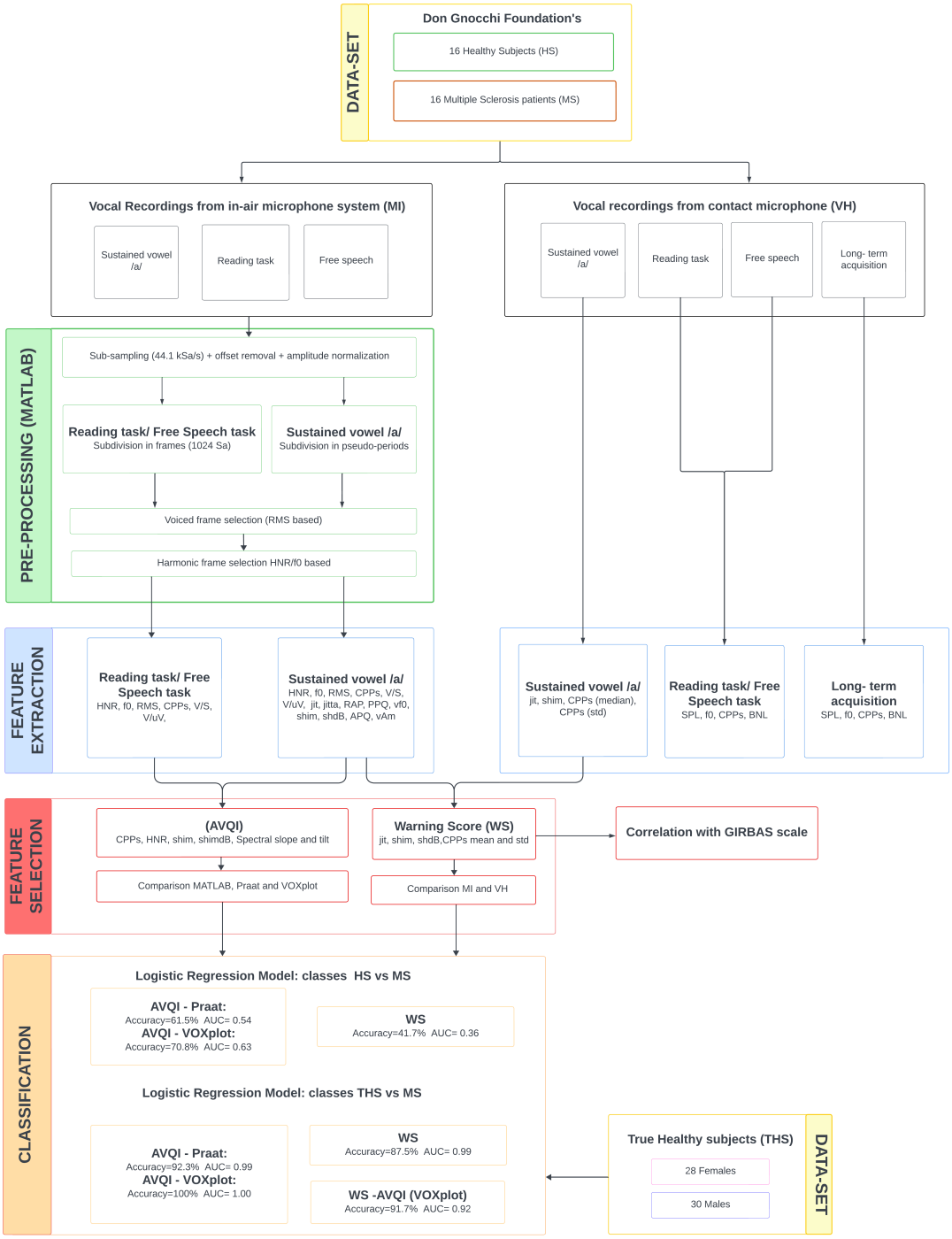

Figure 2.1: Flow-chart showing the various steps performed in this thesis

Figure 2.2: Vocal Holter's kit [10]

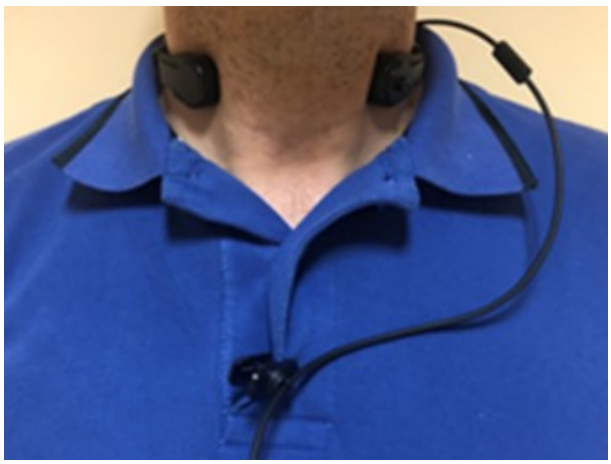

Figure 2.3: Vocal Holter positioning on the subject during acquisition [10]

The voice recordings of the involved subjects, HS and MS, are simultaneously performed with a microphone in air and with a contact microphone-based device. The air microphone system is positioned 30 cm from the mouths of subjects to record the voices of the subjects. It has a resolution of 16 bit, a sampling rate of 44.1 kSa/s and is accessible in .wav format. The skin vibrations brought on by the activity of the vocal folds are measured by the contact microphone system. This device also samples the signal induced by vocal cord activity at a rate of 44.1 kSa/s using 16-bit resolution. As opposed to the in-air microphone, the use

of a contact device allows the effects of sound sources other than the signal of interest to be minimised. The samples acquired with VH are grouped into frames of approximately 46 ms and only the speech frames are processed [11]. Furthermore, it has been reported that the VH device selects the harmonic frames using the same technique as a microphone in the air. To operate the VH device properly, take the following actions:

1. Ensure that the individual is wearing the collar around their neck by connecting the contact microphone to the DAP unit;
2. Switch on the DAP unit and link the computer to the Wi-Fi network it creates;
3. Choose the action to be taken by opening the web interface of the computer;
4. The DAP unit returns the data when the vocalization is finished, and as a result, the estimated parameter value is displayed in a message window. The internal memory stores of DAP unit data that are accessed in the .txt format.

Either short-term or long-term vocal quality assessment can be performed with the VH device:

1. Short-term evaluation: it is used in the continuous emission of the vowel /a/. In case the subject performs more than one vocalization in the total recording time, only the first part will be processed. A message window is shown with the value of several parameters (they will be described in detail later):
   - Fundamental frequency f_0 (expressed in hertz);
   - Local jitter (expressed in percent);
   - Local shimmer (expressed in percent);
   - Median of CPPS (Cepstral Peak Prominence Smoothed) (expressed in decibels);
   - Standard deviation of the CPPS (expressed in decibels).

   Data are returned in the form shown in Figure 2.4, where there are the same five parameters with in addition information pertaining to date and time of monitoring, battery charge, temperature (°C) and relative humidity (%RH).

```
Short evaluation started on date: 2023-02-09 09:32:47
Calibration parameters: Mean squared error: 3.09 dB, linearity: 89.2 %, Intensity ratio: 0.87 : 1
Battery charge: 100 (%) Temperature: 26.2 (C) Relative humidity: 27 (%RH)


Fundamental frequency: 159.8 Hz
Jitter: 2.88 %
Shimmer: 5.48 %
CPPS (median): 18.0 dB
CPPS (standard deviation): 3.98 dB
```

Figure 2.4: Example of short-term evaluation data

2. Long-term evaluation: it was employed in lecture monitoring as well as baseline monitoring, or few minutes monitoring in which the subject was requested to give a brief discussion or read passages from a book. The contact microphone in this monitoring should be fixed in place for the duration of the monitoring.

Long evaluation started on date: 2023-02-09 09:41:01
Calibration parameters: Mean squared error: 3.09 dB, linearity: 89.2 %, Intensity ratio: 0.87 : 1
Battery charge: 100 (%) Temperature: 27.6 (C) Relative humidity: 26 (%RH)

Time	Battery(%)	T(C)	RH(%)	BNL_LAF90(dBA)			BNL_LAF75(dBA)		BNL_LAF50(dBA)				BNL_Leq(dBA)	PPT(%)		SPL_mean(dB)		SPL_median(dB)			SPL_5perc(dB)	
	SPL_95perc(dB)			SPL_SDV(dB)	F0_mean(Hz)		F0_median(Hz)		F0_5perc(Hz)	F0_95perc(Hz)			F0_SD(Hz)	CPPS_mean(dB)		CPPS_median(dB)			CPPS_5perc(dB)		
CPPS_95perc(dB)		CPPS_SD(dB)																				
09:42:30	100	28.3	26	50.1	53.7	58.2	63.9	32	69.2	70.0	65.5	72.5	2.1	153.4	164	77	209	44.2	13.7	15.3	5.7	19.6	4.8
09:43:56	100	28.3	26	50.8	59.9	82.6	63.8	22	69.0	69.5	65.5	73.0	2.3	98.8	81	77	177	37.3	11.5	11.1	5.7	18.9	4.3
09:45:02	100	28.3	26	54.5	58	82.2	65.4	33	70.2	71.0	66.5	73.5	2.4	148.7	163	77	194	39.3	13.7	14.7	6.3	19.1	4.1
09:46:17	100	28.9	30	53.8	67.5	67.6	70.3	36	70.5	71.5	66.0	74.0	2.4	133.8	152	77	205	47.1	14.2	15.7	6.2	19.5	4.3
09:47:30	100	28.9	29	58.1	65.3	63.6	67.5	39	70.2	71.0	66.5	74.0	2.3	118.9	92	77	191	46.5	13.5	14.6	6.2	19.1	4.3
09:48:46	100	28.9	26	78.3	59.4	82.2	64	44	69.8	71.0	65.5	73.0	2.3	107.6	90	77	169	35.5	13.7	15.1	5.9	19.3	4.5
09:50:00	100	29.6	26	57.2	61.1	75.7	66.8	48	70.6	71.5	66.5	74.5	2.4	117.0	100	77	186	38.1	14.2	15.7	6.5	19.9	4.4
09:51:13	100	29.6	25	62.5	64.5	68	70.9	38	70.6	71.5	66.0	75.0	2.8	111.6	93	77	189	41.1	14.6	16.0	6.6	20.0	4.4
09:52:29	100	29.6	26	55.6	61.3	66.6	72.2	32	69.7	70.5	65.5	74.0	2.5	127.4	109	77	196	46.1	13.0	13.7	5.8	19.3	4.5
09:53:42	100	30.3	27	56.9	57.9	66.6	69.4	14	68.8	69.5	64.5	74.5	3.2	153.0	163	77	211	44.5	11.7	12.0	5.4	18.4	4.4
09:54:55	100	30.3	25	54.9	55.5	62	64.9	14	69.4	70.5	64.5	74.5	3.2	117.7	102	77	178	39.1	12.3	12.1	5.7	19.2	4.7

Figure 2.5: Example of long- term evaluation data

The data are returned in the format shown in Figure 2.5, where the first two rows provide initial information: in the first row, we have the date and time of start of monitoring, battery charge, temperature (°C), and relative humidity (%RH), and in the second row, we have calibration parameters. The other returned data provide the various estimated parameters that are updated with a time interval of approximately 75 s:

- For background noise (BNL) there are the statistics LAF_{90}, LAF_{75}, LAF_{50}

and L_{eq}, measured in dB;

- For speech activity there are the percent phonation time (Dt %);
- For sound pressure level (SPL) there are the statistics of mean, median, standard deviation, 5^{th} and 95^{th} percentile, also measured in dB;
- For the fundamental frequency (f_0) there are the statistics of mean, median, standard deviation, 5^{th} and 95^{th} percentile, in Hz.

The next expression is used to express the SPL parameter at the default distance d:

$$SPL_{\mathrm{d}} = SPL_{\mathrm{d_0}} + 20 \log_{10} \frac{d_0}{d}$$

where d is the specific distance of interest and d_0 is 22 cm.

2.2 Data-set

The data set provided by Don Carlo Gnocchi Foundation includes voice recordings of 16 multiple sclerosis patients (MS), with mean age 44 years and standard deviation approximately 14 years, and 16 healthy subjects (HS), with mean age 42 years and standard deviation approximately 12 years, in as noise-free environment as possible, using in-air microphone and a contact-microphone (VH). Some of the subjects HS and MS were not used: in fact those with unclean recordings or no Vocal Holter data were discarded.

Tables 2.1 and 2.2 show the HS and MS data-set with their ID, i.e. the identification codes for privacy, gender and age of each subject. In addition, the material provided by the Don Gnocchi Foundation speech therapists for each subject also includes the year of onset of the disease, other parameters relating to the stage and type of disease and the GIRBAS scale perceptual assessments. For each patients and healthy subject there is a set of recordings acquired either for VH and in-air microphone:

1. Three repetitions of the vowel /a/ at a comfortable pitch, level and duration;
2. Approximately one-minute of free-speech;
3. Reading of a phonetically balanced text called "Notturno". In "Notturno", the frequency of occurrence of each phoneme is equal for all of them.

Table 2.1: Data-set of healthy subjects (HS)

ID	GENDER	AGE
AM66	F	57
BC85	F	37
CA68	F	54
CC85	F	37
CS45	F	77
DM51	F	71
FG93	F	X
LU77	M	45
MI67	M	56
MM73	F	71
PM72	M	51
PN64	M	59
PO65	F	57
SA75	M	47
SM85	M	37
VC77	F	45

Table 2.2: Data-set of pathological subjects (MS)

ID	GENDER	AGE
CF70	M	52
CP46	F	76
CR60	M	63
CS71	F	51
DS76	F	46
DV62	F	60
FR51	F	72
FS94	F	28
GF77	M	45
GG72	M	50
LA71	M	51
MC84	M	38
MG77	M	45
NP69	M	53
PM43	F	79
SA49	F	73

There are equally present harmonic and non-harmonic sections in addition to the inevitable silences in the recording caused by gaps between phrases.

"Notturno. Vi è un profondo silenzio nel buio della notte. Vicino 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."

2.3 Pre-processing

For this analysis, it was decided to examine both the free speech and the lecture of the text as well as the three repeats of the vowel /a/ in order to evaluate their vocal abilities completely. However, some records are missing as a result of logistical and organizational issues, therefore it was necessary to divide the data into smaller groups for analysis in order to make the most of what was available. Additionally, due to saturation issues or problems with the quality of

the microphone and acquisition system, several recordings were initially rejected. Following this initial selection, the recordings were examined using the software *Audacity* (version 3.2.5), which made it possible to exclude signal segments that are characterized by instrumental artifacts or outside noise. In order to avoid truncating the voice signal, care was made during hand removal to avoid cutting sections of the recordings during phonation. The cut should start and terminate only during silences.

The *pre-processing* is done simultaneously for the three tasks (free speech, balanced speech, and sustained vowel /a/). Using the software Audacity (version 3.2.5) to help, all the records are first listened to and examined in order to eliminate the voice pieces' opening and ending sections, which are deemed unnecessary for the purpose of this work. Next, in the Matlab (R2022b) environment, each signal (corresponding to a single subject) is loaded and re-sampled at 44.1 kSa/s, a value that is frequently utilized in literature. Subsequently, the mean value for the complete signal is controlled. It is eliminated if the mean value exceeds 20% of the root mean square RMS value. Normalization with regard to amplitude is the next phase, when the signal is adjusted to the highest absolute value of the signal under analysis. Following this phase, a subdivision is carried out in accordance with the specific job being analyzed. In the case of balanced and free speech, signal samples are grouped into frames of 1024 samples, or a time interval of 23 ms, which is equivalent to the inter-syllabic pauses in terms of magnitude. Instead in the case of the vowel's three repetitions, signal samples in /a/ are categorized into pseudo-periods, which are distinguished by an algorithm for auto-correlation. The removal of silent frames is a crucial operation that is made possible by the employment of a predetermined threshold that is half the RMS value of the entire signal. More specifically, 1024-sample frames are moved over the signal, and if the RMS value of any frame above the predetermined threshold, it is regarded as a voiced, non-silence frame and is stored in an array for further processing. In the event that the outcome is negative, it might be discarded as a quiet frame. The next control is used to choose harmonic frames based on a predetermined standard. This criterion is to extract the parameters HNR and f_0 value from each voiced frame (the process of determining these two measures will be covered later in this chapter). Then, only the frames with an HNR value greater than 0 dB and a frequency jump between adjacent frames that is neither lower than -25% nor

higher than +50% are chosen. Because of this check, only frames with harmonic content that is at least as high as the noise energy in each frame will move on to the next section of the research, which is the feature extraction phase.

2.4 Feature Extraction

After pre-processing, only the signal blocks that were selected for *feature extraction* were used. In the case of free speech and reading task, the signal is observed using windows with a fixed number of samples (frames of 1024 samples). Instead in the case of the vowel /a/, being an almost periodic signal, it is better to use hypothetical periods as frames. Two different scripts were used for feature extraction, one for the vowel /a/ and one for the reading and free speech task. The auto-correlation of the signal, which theoretically reaches a local maximum at the shift value equal to the period of the signal, i.e. where the signal repeats more or less the same, was used to calculate the pseudo-periods. The absolute maximum is at the zero shift, which corresponds to the power of signal. The auto-correlation method is also used to determine the fundamental frequency, HNR, and pseudo periods. The steps of the process are briefly detailed in the section that follows. RMS, HNR, F_0, and CPPS were the four acoustic characteristics that were used to evaluate and contrast the patient's vocal performances. The parameters were evaluated using their respective statistical distributions, taking into account the following statistics for each one: mean, median, and mode as the central trend; standard deviation, range, 5° percentile, and 95° percentile as the variability measure; and skewness and kurtosis as the shape factor. Nine stability parameters in terms of time and amplitude are also calculated for the vowel /a/. The software instruction manual MDVP (Model 5105) [12] was used for the implementation of some parameters in *Matlab* (R2022b) summarized in Figures 2.3, for reading and free speech task, and Figure 2.4, for sustained vocal /a/. The definitions of the recorded metrics and acoustic parameters are provided in paragraph 2.5.

Table 2.3: Extracted features for balanced text reading and free speech task

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

Table 2.4: Extracted features for sustained vowel /a/ 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, PPS (%)
Coefficient of Foundamental Frequency Variation, vf_0 (%)
Shimmer, ShdB (dB)
Shimmer Percent, Shim (%)
Amplitude Perturbation Quotient, APQ (%)
Coefficient of Amplitude Variation, vAm (%)

2.5 Parameters

The parameters of interest described below were extracted from the voice signal acquired by the microphone in the air.

2.5.1 Acoustic Parameters

- *RMS - Root Mean Square*
  In zero mean alternating signals, the RMS value is utilized as a power indicator. For the calculation, Matlab's built-in "rms" function was utilized.

$$x_{\mathrm{RMS}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x_n^2} \tag{2.1}$$

  where x_n are samples of the signal. As previously mentioned, the RMS value of each frame serves as a discriminator for the quiet frames and is determined prior to the extraction of the other acoustic data. The non-silent frames' RMS value is then stored in an array.

- *HNR - Harmonic to Noise Ratio*
  The harmonic-to-noise ratio is a parameter for quantify the harmonicity of the signal and can be calculated from autocorrelation.
  In signal theory, the autocorrelation function Rx[n] in the discrete is defined as:

$$R_{\mathrm{x}}[n] = \sum_{k=-\inf}^{+\inf} x^*[k]x[k+n] \tag{2.2}$$

  where n is the lag (delay). In the autocorrelation function we have the maximum in the origin, which coincides with the average power of the signal. The function of normalized autocorrelation $R'_{\mathrm{x}}[n]$:

$$R'_x[n] = \frac{R_{\mathrm{x}}[n]}{R_{\mathrm{x}}[0]} \tag{2.3}$$

  Consider the additive noise model in which we assume that we have a periodic signal *h[n]* to which noise *N[n]* is added: *x[n] = h[n]+N[n]*. The autocorrelation function normalized to delay *[n] = [τ_{max}]* represents the relative power of the periodic (or harmonic) component of the signal, and its complement represents the relative power of the component of noise [13]:

$$R'_x[\tau_{max}] = \frac{R_{\mathrm{H}}[0]}{R_{\mathrm{x}}[0]}; 1 - R'_x[\tau_{max}] = \frac{R_{\mathrm{N}}[0]}{R_{\mathrm{x}}[0]} \tag{2.4}$$

  Harmonic-to-noise ratio (HNR) is defined as:

$$HNR = 10log\frac{R'_x[\tau_{max}]}{1 - R'_x[\tau_{max}]} \tag{2.5}$$

It is important to note that whereas low values of HNR indicate a noise component higher than the periodic, high values of HNR imply a periodic component in the signal greater than the noise component. The harmonic-to-noise ratio has an infinite value for precisely periodic signals, which is the limiting case. Through the use of the logarithmic operation, HNR is measured in dB; when it is larger than zero, it indicates that the periodic signal power is greater than the non-periodic signal power. Harmonicity of non-silent frames is determined by calculating their HNR value: if the value exceeds 0 dB, the frame is selected, the HNR value is kept in an array, and the extraction proceeded; if the value of HNR is not higher than 0 dB, the frame is deemed non-harmonic. This is because for healthy voices the HNR value is in most cases above 0 dB, on the other hand, pathological voices can sometimes lead to negative HNR values, i.e. the energy of the harmonic component is lower than the noise level.

- *F_0 - Foundamental Frequency*
  The fundamental frequency F_0 is another variable that is dependent on autocorrelation. This value is crucial because it may be used to estimate parameters that affect the stability of period and amplitude.
  The autocorrelation method [13] was used to determine the length of a pseudo-period T_0 (or of the fundamental frequency $F_0 = \frac{1}{T_0}$). Both speech (free speaking or reading a passage) and sustained vowels make sense according to these criteria.

- *CPPS - Cepstral Peak Prominence Smoothed*
  For the complete understanding of the parameter, it is necessary to briefly explain what the Cepstrum is and how it is obtained.The concept of cepstrum was introduced in 1963 by Bogert et al. and the name was obtained by anagramming the word "spectrum".
  The cepstrum can be defined as a spectrum of a logarithmic spectrum: if the first spectrum (power spectrum) indicates the energy of the signal around each frequency, the second spectrum indicates how periodic the harmonic components in the spectrum. Mathematically, the cepstrum $C(\tau)$ is defined

as:

$$C(\tau) = |\mathcal{F} log(|\mathcal{F}\{f(t)|^2)\}|^2 \tag{2.6}$$

where $\mathcal{F}$ is the Fourier transform, $|\mathcal{F}\{f(t)|^2$ is the power spectrum, and $f(t)$ is the signal as a function of time. The variable τ is called "quefrency" (a word derived from the anagram of the word "frequency") and has the dimensions of time. Cepstral analysis is a technique that is used in several fields, includingin speech analysis.

Mathematically, the source-filter system of the phonatory system can be schematized by the following relationship:

$$S(z) = H(z) \cdot U(z) \tag{2.7}$$

where $H(z)$ is the transfer function of the vocal tract, $U(z)$ is the spectrum of the excitation signal, i.e., the spectrum of the signal arriving from the chords vocal tract, and $S(z)$ is the spectrum of the resulting sound.

By the properties of the convolution theorem, in the time domain, equation 2.7 becomes:

$$s(z) = h(z) * u(z) \tag{2.8}$$

According to the logarithm property, the logarithm of a pair of numbers is equal to the sum of their individual logarithms. Then, in the cepstrum's domain, equation 2.7 becomes:

$$log|S(z)| = log|H(z)| + log|U(z)| \tag{2.9}$$

In other words, the fundamental period component and the vocal tract have multiplicatively correlated relationships in the frequency domain, whereas they have additive relationships in the cepstrum domain.

The basic period is represented by the cepstral peak, which is distinct from the vocal tract component. It is possible to gauge the voice's pitch by using the cepstrum in speech analysis.

In reality, the resonance frequency of the peak cepstral is the average fundamental frequency of the frame under consideration, and the prominence in relation to the noise floor says about the harmonic nature of the signal. Additionally, the cepstrum can be used to gather spectral envelope data for

speech analysis.
The CPP (Cepstral Peak Prominence) and the CPPS (CPP-smoothed) are the parameters that can be derived from the cepstrum [14]. The latter enables accurate estimation of the prominence of the cepstral peak using techniques that smooth background noise to lessen its impact on peak measurement. The CPP (Cepstral Peak Prominence), which is expressed in decibels (dB), is the difference between the cepstral's amplitude and the peak value of the regression line [15].

The CPPS (CPP-Smoothed) measures CPP on a mediated version of the cepstrum. There are two phases to smoothing [14]. The cepstral are averaged over time in the first phase; the current cepstrum is averaged with a number of cepstrums that come before and after the cepstrum under consideration. The cepstrum are averaged along the frequency in the following step: by averaging the cepstrum values over number of bins. The parameter of interest is measured on the later cepstrum. The location and magnitude of the cepstral peak must be known in order to calculate the CPPS. A cepstrum interval roughly ranging from 3.3 ms to 16.7 ms, which corresponds to the frequency range from 60 Hz to 300 Hz, is used to measure these two parameters. In fact, the human voice's basic frequency, F_0, can roughly lie inside this range.
The voice signal was subsampled by a factor of 2 in order to measure the cepstral parameter. From there, the CPPS was calculated from frames that were 1024 samples long (46.4 ms) and overlapped by 44 samples (2 ms). The measured signal can be sufficiently covered in this manner.
The cepstral characteristics have an important role in reading and speech.

2.5.2 Recording Parameters

The percentage of frames rejected in the two pre-evaluation processes prior to feature extraction is represented by three pre-processing output parameters that are also recorded; they are helpful for interpreting the other parameters. The length of the signal under investigation, in particular, has a significant impact on the stability parameters that follow; in fact, because they are the outcome of an averaging procedure, the longer the signal, the parameters will become more stable.

Here are the calculations made:

- *Non-silent frame ratio (%)*:

$$\frac{V}{S} = 100 \cdot \frac{n_{voiced}}{n_{voiced} + n_{unvoiced}} \tag{2.10}$$

  where the number of voiced frames n_{voiced} includes harmonic and non-harmonic frames and $n_{unvoiced}$ indicates the number of silent frames;

- *Harmonic frames ratio (%)*:

$$\frac{V}{uV} = 100 \cdot \frac{n_{harmonic}}{n_{harmonic} + n_{non-harmonic}} \tag{2.11}$$

- *Length*: the quantity of valid frames following pre-processing.

2.5.3 Stability Parameters

With the exception of nine additional parameters of amplitude and period stability for the vowel /a/ analysis, the parameters utilized for the analysis of the three repeats of the vowel /a/ and free speech are the same. In fact, using these characteristics can help determine how much the patient can maintain a consistent voice and tone. Below are the 9 parameters and their accompanying definitions:

1. Jita [μs] : *Absolute Jitter*
   A measurement of the pitch period's period-to-period fluctuation. Areas where voices break up are excluded.

$$Jita\ [\mu s] = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}| \tag{2.12}$$

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

2. Jitt [%]: *Jitter Percent*
   Relative estimation of the variation from time to period.

$$Jitt\ [\%] = 100 \cdot \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}|}{\frac{1}{N} \sum_{i=1}^{N} T_0^{(i)}} \tag{2.13}$$

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

The same kind of pitch interruption is assessed by both Jitta and Jitt. The failure of vocal cords to maintain a periodic vibration for a specific amount of time is a possible cause of periodic irregularity. These changes are typically random. Hoarse voices are frequently connected to them. Jitta is an absolute measurement that depends on typical fundamental frequency of the voice. Because of this, the normative ideals from the Jitta for men and women are very different. Jitta is challenging to compare because lower Jitta is correlated with higher pitch and vice versa. In contrast, because Jitter is a relative measurement, the average fundamental frequency of the topic has a far smaller impact.

3. RAP [%] : *Relative Average Perturbation*

   A three-period smoothing factor was used to evaluate the period-to-period fluctuation of the pitch within the studied voice sample.

$$RAP\ [\%] = 100 \cdot \frac{\frac{1}{N-2}\sum_{i=2}^{N-1} |\frac{T_0^{(i-1)}+T_0^{(i)}+T_0^{(i+1)}}{3} - T_0^{(i+1)}|}{\frac{1}{N}\sum_{i=1}^{N} T_0^{(i)}} \tag{2.14}$$

   where: $T_0^{(i)}$, i = 1,2.. N extracted pitch period data, N: number of extracted pitch periods. It is comparable to the Jitt but smoothed such that RAP is less sensitive to faults in pitch extraction. It is, however, less sensitive to relatively brief fluctuations. RAPs may be higher in voices that are hoarse or breathy.

4. PPQ [%]: *Pitch Period Perturbation Quotient*

   Relative assessment of the period-to-period of the pitch fluctuation within the studied voice sample using a 5 period smoothing factor.

$$PPQ\ [\%] = 100 \cdot \frac{\frac{1}{N-4}\sum_{i=1}^{N-4} |\frac{1}{5}\sum_{r=0}^{4} T_0^{(i+r)} - T_0^{(i+2)}|}{\frac{1}{N}\sum_{i=1}^{N} T_0^{(i)}} \tag{2.15}$$

   where: $T_0^{(i)}$, i = 1,2...N extracted pitch period data, N: number of extracted pitch periods Similar to RAP in many ways, PPQ features smoothing over five

periods rather than three, therefore the effect of smoothing is more pronounced.

5. vf_0 [%]: *Coefficient of Foundamental Frequency Variation*
   Relative standard deviation of the fundamental frequency. It reflects, generally, the range of f_0 (short to long-term) in the voice sample under analysis. The standard deviation of the extracted period-to-period ratio is used to calculate vf_0 by the average fundamental frequency as follows:

$$vf_0\ [\%] = 100 \cdot \frac{\sigma}{f_0} = 100 \cdot \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N}\sum_{j=1}^{N} f_0^{(j)} - f_0^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N} f_0^{(i)}} \tag{2.16}$$

   where: $\mathrm{f}_0 = \frac{1}{N}\sum_{i=1}^{N} f_0^{(i)}$ and $f_0^{(i)} = \frac{1}{T_0^{(i)}}$ period to period fundamental frequency values and $T_0^{(i)}$, i= 1,2 .. N extracted pitch period data, N: number of extracted pitch periods. vf_0 reveals changes in the fundamental frequency. Regardless of the type of pitch variation, the vf_0 value rises. The vf_0 value rises whether the variations are brief or long-term, random or predictable. These fluctuations can simply be an increase or decrease in pitch, frequency tremors, non-periodic variations, or both.

6. Shim [%]: *Shimmer Percen*
   Relative assessment of the peak-to-peak amplitude variability within the studied voice sample from period to period (very short term).

$$Shim\ [\%] = 100 \cdot \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A^{(i)} - A^{(i+1)}|}{\frac{1}{N}\sum_{i=1}^{N} A^{(i)}} \tag{2.17}$$

   where: $A^{(i)}$, $i = 1, 2...$N extracted peak to peak amplitude data, N: number of extracted impulses. The inability of the strings to sustain a periodic vibration for a specific period and the existence of turbulent noise may be linked to cycle-to-cycle amplitude inconsistency. Hoarse and breathy voices are generally related to this kind of random irregularity.

7. ShdB [*dB*]: *Shimmer in dB*
   Evaluation in dB of the period-to-period varibiality of the peak-to-peak amplitude within the analyzed voice sample.

$$ShdB\ [dB] = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(\frac{A^{(i+1)}}{A^{(i)}})| \tag{2.18}$$

   where: $A^{(i)}$, $i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. Shim and ShdB employ two distinct measures, % and dB, but both are relative assessments of the same class of amplitude perturbation.

8. APQ [%]: *Amplitude Perturbation Quotient*
   Relative assessment of the peak-to-peak amplitude period-to-period variability within the studied voice sample at a smoothing of 11 periods.

$$APQ\ [\%] = 100 \cdot \frac{\frac{1}{N-10}\sum_{i=1}^{N-10}|\frac{1}{11}\sum_{r=0}^{10}A^{(i+r)} - A^{(i+5)}|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.19}$$

   where: $A^{(i)}$, i= 1, 2...N extracted peak to peak amplitude data, N: number of extracted impulses. A parameter called APQ is very comparable to shimmer, however it has a smoothing factor of 11. Although smoothing lessens sensitivity of APQ to amplitude variations from one period to the next, it still does a great job of describing short-term amplitude perturbations of the voice. Voices that are wheezy and hoarse typically have higher APQ.

9. *vAm* [%]: *Coefficient of Amplitude Variation*
   Relative standard deviation of the peak-to-peak amplitude. It generally reflects the short- to long-term peak-to-peak amplitude changes found in the examined voice sample.

$$vAm\ [\%] = 100 \cdot \frac{\sigma}{A_0} = 100 \cdot \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N}\sum_{j=1}^{N}A^{(j)} - A^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.20}$$

   where: $A^{(i)}$, $i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. Cycle-to-cycle amplitude variation of the voice are revealed by vAm. Any change, whether short-term or long-term, predictable or arbitrary, raises the value of vAm.

As a result, matrices are created that display the extracted parameters in the columns and the time observations of the individual patients under analysis in the rows. 39 features are reported for the free speech on the columns, compared to a total of 47 features for the three /a/ repetitions.

2.6 Feature Selection

Once the parameters were extracted, the selection of the characteristics was carried out using two indices: *Acoustic Voice Quality Index* (AVQI), known from the literature, and another new one, developed for the assessment of vocal behaviour, called *Warning Score* (WS).

2.6.1 Acoustic Voice Quality Index

Vocal quality plays a crucial role in human communication and significantly influences the perception of the effectiveness and expressiveness of vocal messages. Over the years, numerous tools have developed for objectively assessing vocal quality: one is the *Acoustic Voice Quality Index* (AVQI), which represents a significant advancement in the objective measurement of vocal quality. In this section, AVQI will be explained in detail, including its calculation methods, clinical applications, and future prospects in voice research.
Acoustic measures offer an unbiased assessment of voice quality, serve as a non-invasive and practical method for quantifying dysphonia, and provide evidence of functional improvements following medical, surgical, and rehabilitation treatments. Since recording and processing sustained vowels is quick and easy, sustained vowels are the focus of acoustic analysis. Furthermore, phonetic, prosodic, and linguistic disturbances have a less impact on steady state phonations [16]. Continuous speech samples, on the other hand, are more typical of daily speech use and frequently show symptoms of voice issues better than sustained vowels [17].
AVQI is an index first developed by Maryn [18] as a tool for objectively measuring vocal quality. This index was conceived to overcome the limitations of subjective qualitative assesments by providing a quantitative measure based on acoustic data. AVQI is a complex construct built on the foundation of linear regression analysis

that combines a number of acoustic factors to produce a single score for the evaluation of overall voice quality.

There are different factors that have contributed to this index [19]:

1. Continuous speech entails quick, frequent shifts brought on by glottal and supraglottal mechanisms, whereas a prolonged vowel indicates relatively time-invariant phonation.

2. Sustained mid-vowel segments lack prosodic basic frequency and amplitude changes, quick voice onsets and terminations, and non-voiced phonemes.

3. Long vowels speak at a velocity that is unaffected by vocal pauses, phonetic context, and tension.

4. The standard fundamental time or frequency measures of perturbation and amplitude perturbation heavily rely on algorithms for pitch identification and extraction. They become less accurate when speaking continuously.

5. It is possible to elicit and produce sustained vowels with less work and in a more consistent way.

AVQI has a wide range of clinical applications that make it a valuable tool in various contexts:

1. Vocal Diagnosis: AVQI can assist in identifying and assessing vocal abnormalities, providing medical professionals and vocal specialists with objective information about an individual's vocal quality. This is particularly useful in early diagnosis of pathological vocal conditions [20].

2. Therapeutic monitoring: during vocal therapy, AVQI can be used to monitor a patient's progress over time. This allows vocal specialists to tailor treatments more precisely and evaluate the effectiveness of therapies [21].

3. Vocal research: AVQI opens up new avenues for systematically and objectively studying vocal quality. AVQI-based studies can contribute to the understanding of vocal dynamics in various contexts, from phonetics to vocal psychology [22].

The utility of AVQI in clinical applications and vocal research makes it a growing field of interest. Ongoing research and development in the reign of AVQI promise to enhance the diagnosis and management of vocal conditions and advance our understanding of the human voice. By using the presumptions of the Finnish language, Fantini and his team tried to verify an AVQI for the Italian language [23]. The study included both euphonic and dysphonic subjects. Patients with dysphonia might have both organic and inorganic diseases, and the severity of their dysphonia can vary.

The AVQI uses concatened 3 seconds of a sustained vowel /a/ and voice segments from a phonetically balanced text and consists of weighted combination of time, frequency and quefrency-domain metrics, as seen in 2.5 paragraph.

$$\begin{aligned}\text{AVQI} = [\ 4.152 - (\ 0.177 * \text{CPPS}\ ) - (\ 0.006 * \text{HNR}\ ) - (\ 0.037 * \text{Shimm}\ ) + \\ (0.941 * \text{ShdB}\ ) + (\ 0.01 * \text{Slope}\ ) + (\ 0.093 * \text{Tilt}\ )\ ] * 2.8902\end{aligned}$$

The recording of the sustained vocal /a/ and reading task was carried out using the same equipment, under standard conditions and with the microphone held at a distance of one metre from the mouth of the subject. In this study, Matlab's results were compared with two software applications: Praat and VOXplot.

- **Praat** is a widely used open source software for language analysis and synthesis and sound processing. It was mainly developed by Paul Boersma and David Weenink at the University of Amsterdam [24]. Praat is commonly used in fields such as linguistics, phonetics, psycholinguistics and musicology to conduct sound analysis and language studies. Praat's main functionalities include:
  - Sound analysis: Praat allows users to record, visualise and analyse audio waveforms. It can be used to perform measurements on sound signals, such as fundamental frequency, intensity and more.
  - Phonetic analysis: The software can perform advanced phonetic analysis, including spectral analysis, formant and phonetic segmentation.
  - Speech synthesis: Praat can be used to generate speech synthesis, allowing users to create synthetic voices from phonetic data.
  - Sound manipulation: Users can modify sound in various ways, for example to remove background noise or to perform pitch analysis.

– Phonetic transcription: Praat supports phonetic transcription, allowing users to tag and annotate segments of the sound according to their phonetic perception.

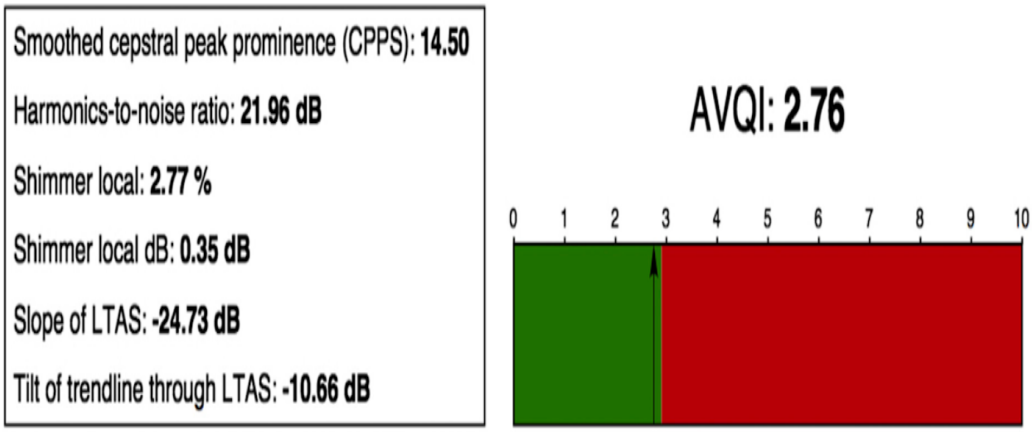

Figure 2.6: Example of the outcome of the main voice quality parameters in Praat

Praat is a very versatile tool that is widely used by researchers and scholars in the fields of linguistics and phonetics to perform detailed sound and language analysis. Using the script, which can be found in Appendix A, one can also obtain an analysis of certain parameters such as CPPS, HNR, local Shimmer in % and dB, Spectral Slope and Tilt. In addition, Figure 2.6 shows an example of the outcome of the voice quality parameters in Praat.

- **VOXplot** is a new freeware program that analyzes the acoustic voice quality using the Praat signal processing methods. VOXplot is designed primarily for speech quality analysis, while Praat is a versatile and equally complex software for acoustic analysis of arbitrary signals. Whereas user interface of VOXplot is designed to meet the requirements in terms of standardised and intuitive user-friendliness, Praat exclusively uses algorithms. The complete acoustic voice quality evaluation workflow is covered by VOXplot, including recording and quality assessment, acoustic voice quality analysis, and the creation of a brief PDF sheet with the analysis results, as it can be seen in Figure 2.7. [25]

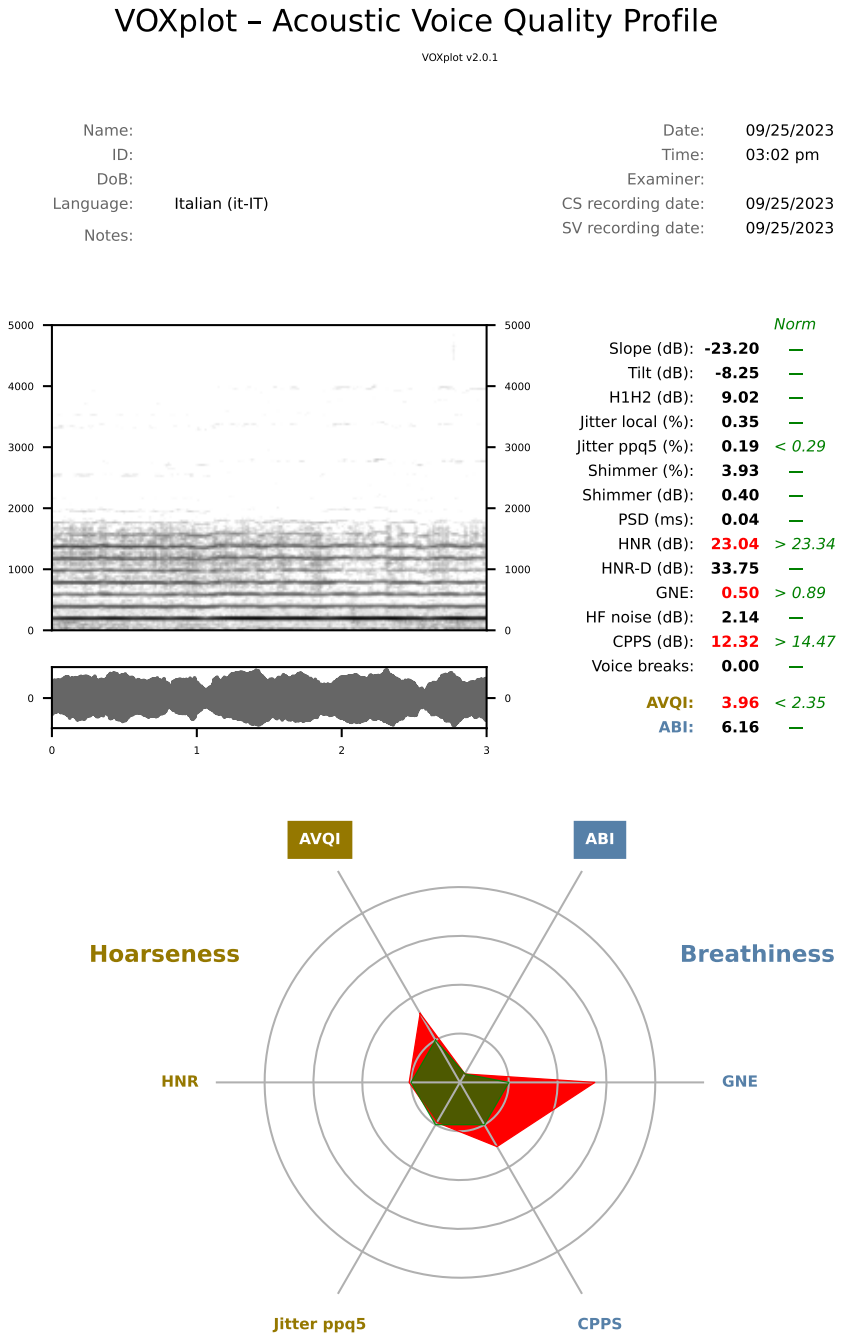

Figure 2.7: Example of the outcome of the main voice quality parameters in VOXplot, which are evaluated quantitatively and/or qualitatively for hoarseness and breathiness for the healthy subject AM66

2.6.2 Warning Score

A new index called *Warning Score* [26] was used to assess the vocal health status of the subjects and was assigned for each subject from the parameters extracted from the VH and the microphone in the air. As mentioned before, for each subject, both pathological and healthy one, three vocal recording was done with the device Vocal Holter (VH) and Microphone in-air (MI):

1. Sustained vowel /a/ repeated 3 times;

2. Reading of a phonetically balanced text;

3. Free speech of about one minute.

In particular, from the repetition of the sustained vowel /a/, *local jitter* (%), *local shimmer* (%), *mean* and *standard deviation* of the *Cepstral Peak Prominence smoothed* CPPS (dB) were extracted. Voice recordings are classified as healthy, pathological or "not reliable" in dependence on the values of these parameters. Therefore WS is given according to the following procedures:

- A value less than 0.31% suggests a healthy voice, whereas a number greater than 0.43% suggests a pathological voice. If the local jitter value is between (0.31 - 0.43)%, it is considered unreliable.

- A local shimmer value of less than 2.37% indicates a healthy voice, more than 2.55% indicates a pathological voice, and if it varies between (2.37 - 2.55)%, it is deemed unreliable.

- The $\mathrm{CPPS_{mean}}$ value is considered unreliable if it is between (18.0 - 19.7) dB, a value less than 18.0 dB indicates a pathological voice, and value higher than 19.7 dB indicates a healthy voice.

- A $\mathrm{CPPS_{std}}$ value of less than 0.9 dB indicates a healthy voice, greater than 1.3 dB indicates a pathological voice, and if it falls between (0.9 - 1.3) dB, it is deemed unreliable.

2.7 Logistic Regression

A fundamental statistical technique used for binary and multiple-class classification applications is *logistic regression* (LR). It is a type of statistical technique that examines the correlation between a group of independent variables and a set of binary dependent variables. It is an effective instrument for making decisions. Regression uses a sigmoid function to estimate the probability for the given class using the output of a linear regression function as input. Logistic regression differs from linear regression in that it predicts the likelihood that an instance will belong to a specific class or not, whereas the output of the former is a continuous value that can be anything. Logistic regression is used for predicting the categorical dependent variable using a given set of independent variables. In particular, linear regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems. The logistic function, commonly referred to as the sigmoid function, is the fundamental component of logistic regression and converts input values into a range between 0 and 1. It can be either Yes or No, 0 or 1, True or False but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. The logistic function is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{2.21}$$

Where z is a linear mixture of the weights assigned to the input features:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_\mathrm{p} x_\mathrm{p} \tag{2.22}$$

β_0, β_1, ..., β_p are the model parameters (coefficients) that are learned during the training process, and x_1,x_2,...,x_p are the input features.

Logistic regression models are used to estimate the probability of an event occurring. To provide a binary classification, the probability is compared to a predetermined threshold, usually equal to 0.5. Given the logistic function, the estimated probability of the positive class *(P(Y=1))* can be expressed as:

$$P(Y = 1|X) = \sigma(z) \tag{2.23}$$

Where X represents the input features.
Optimizing the likelihood function during model training in logistic regression often requires employing methods like gradient descent. The goal is to determine the

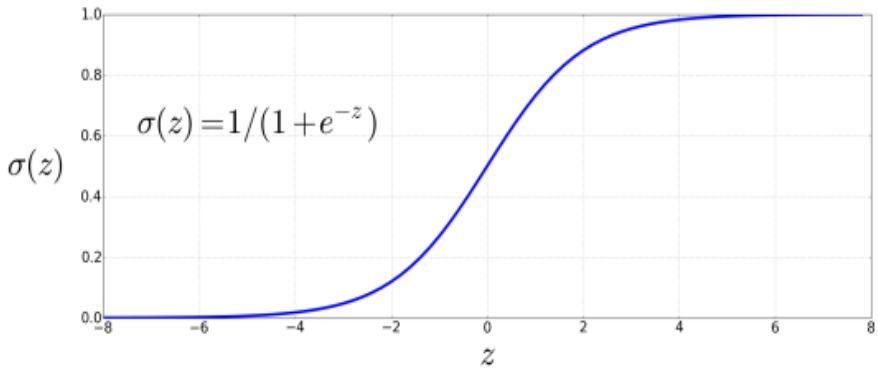

Figure 2.8: The sigmoid function σ(z) takes a real value and maps it to the range(0,1).It is nearly linear around 0 but outlier values get squashed toward 0 or 1

ideal values of β_0, β_1, ..., β_p that maximize the likelihood of the observed data.

For binary classification tasks, binary logistic regression is employed. Each observation in the logistic regression model is classified into one of two groups (for instance, 0 or 1) as it can be seen in Figure 2.8. The coefficients β provide information on how the input attributes and the event's log-odds relate to one another. Numerous metrics, including accuracy, precision, recall, F1-score, and the ROC curve, are frequently used to evaluate the performance of a binary logistic regression model.

After obtaining the prediction results, it is possible to assess performance by contrasting the predicted results of algorithm with the actual results in a *confusion matrix* (CM). Confusion matrices, sometimes referred to as error matrices or contingency tables, are particular table arrangements that make it possible to visualize performance of a classifier under supervised training. The Confusion Matrix, in Figure 2.9, compares 4 result categories whose sum must return the total number of observations. In reality, if one considers the binary encoding of 0 and 1 as Negative and Positive, respectively, the algorithm can behave in 4 different ways with the binary data:

- True Positives (TP): number of correctly classified positive values;
- True Negatives (TN): number of correctly classified negative values;
- False Positives (FP): number of negative elements misclassified (i.e., assigned

to the negative class);

- False Negatives (FN): number of positive elements misclassified (i.e., assigned to the negative class).

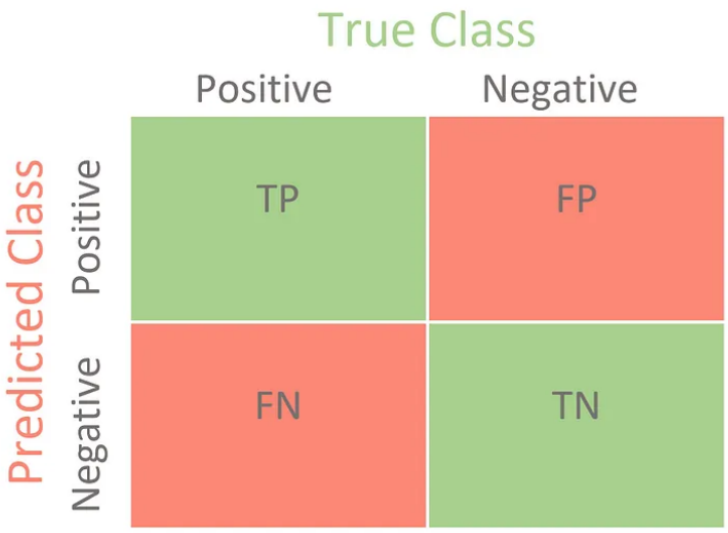

Figure 2.9: Confusion Matrix for Binary Classification

These quantities are helpful in analyzing metrics that measure the performance of classifier. The most popular metrics include:

- Accuracy [%]: it is the most intuitive and widely used metric for evaluating classification models. It measures the percentage of correct predictions compared to the total number of predictions made. The formula is as follows:

$$ACC = 100 \cdot \frac{TP + TN}{TP + TN + FP + FN} \tag{2.24}$$

- Precision [%]: it measures the quality of positive predictions made by a classification model. It is particularly important in situations where making false positive predictions is costly or undesirable. In summary, precision indicates how accurate the model is in its positive predictions.The formula is as follows:

$$PRE = 100 \cdot \frac{TP}{TP + FP} \tag{2.25}$$

- Sensitivity or Recall [%]: it measures the ability of the model to identify all positive examples. It is particularly important in problems where false negatives are costly or dangerous. The formula is as follows:

$$SENS = 100 \cdot \frac{TP}{TP + FN} = TPR = 1 - FNR \tag{2.26}$$

- F1-score: it is a Precision-Recall Trade-off. Increasing precision may decrease recall and vice versa. The F1-score is a metric that combines precision and recall into a single measure:

$$F1 - score = \frac{2 \cdot PRE \cdot SENS}{PRE + SENS} \tag{2.27}$$

- Specificity [%]: is a metric that measures a classification model's ability to correctly identify negative instances. Specificity is complementary to sensitivity (recall), which measures the ability of the model to detect all positive instances. The formula is as follows:

$$SPEC = 100 \cdot \frac{TN}{TN + FP} = TNR = 1 - FPR \tag{2.28}$$

- ROC Curve (Receiver Operating Characteristic): it is a graphical representation of the relationship between sensitivity (recall) and specificity of a model as the classification threshold varies. Sensitivity (true positive rate) is plotted on the y-axis, while specificity (true negative rate) is plotted on the x-axis. To construct an ROC Curve, the model is evaluated at various classification thresholds. At each threshold, the true positive rate (TPR) and false positive rate (FPR) are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{2.29}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.30}$$

  These TPR and FPR values are then used to plot the ROC Curve. As the classification threshold changes, a series of points on the ROC Curve is obtained, representing the ability of the model to distinguish between positive and negative classes [27].

- AUC (Area Under the ROC Curve): it is a numerical metric that measures the overall effectiveness of a classification model. AUC calculates the area under the ROC Curve, thus providing a single value that reflects the discrimination capability of model (Figure 2.11) . AUC ranges from 0 to 1, where a value of 0 indicates poor discrimination ability (worse than a random model), and a value of 1 indicates perfect discrimination capability as it can be seen in Figure 2.10.
  - An AUC of 0.5 indicates that the discrimination ability of the model is no better than that of a random model.
  - An AUC greater than 0.5 indicates that the model has better discrimination ability than a random model.
  - An AUC of 1 represents perfect discrimination, where the model completely separates positive and negative classes.

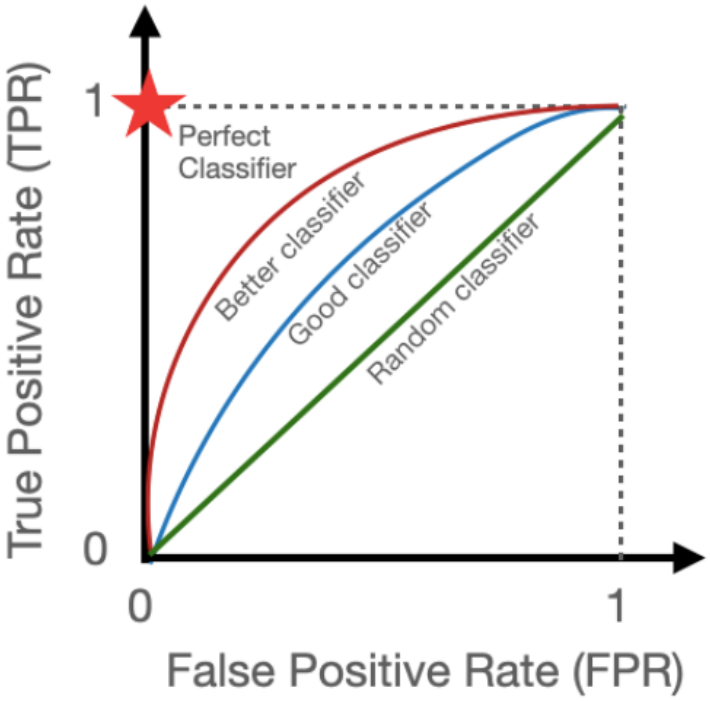

Figure 2.10: Examples of ROC curves of different classifiers [28]

The ROC Curve (Receiver Operating Characteristic) and the Area Under the

Curve (AUC) are fundamental evaluation metrics used to measure the performance of classification models, especially in binary problems. These metrics provide a detailed analysis of discrimination capability of the a model and are widely employed in fields such as medicine, security, marketing, and machine learning.

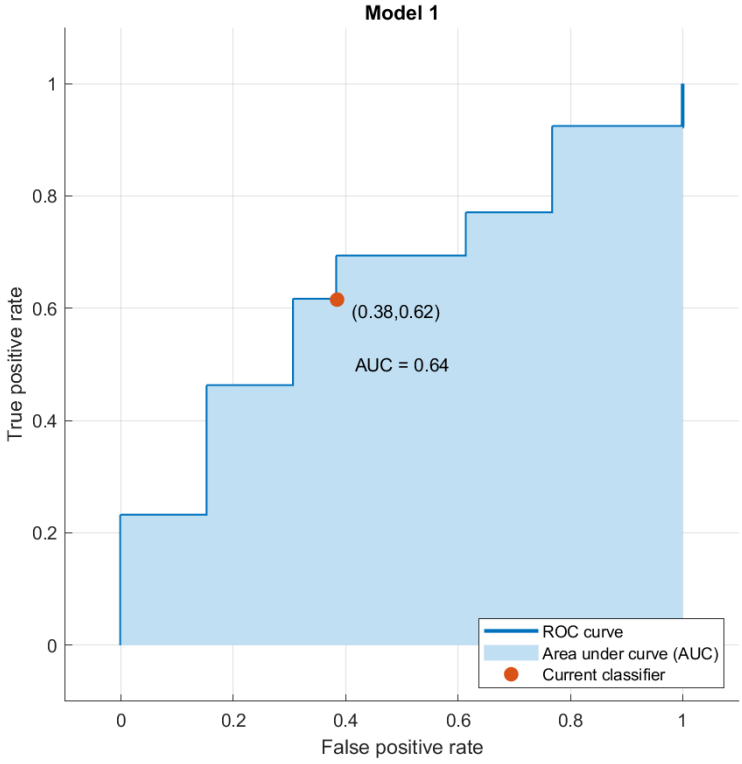

Figure 2.11: Example of ROC curve computed by the Classification Learner App in Matlab (R2022b)

2.7.1 Classification using Logistic Regression

Utilizing the Classification Learner included in Matlab (R2022b) APPs, the subjects were classified according to AVQI and WS. This interface enables manual selection of the subjects to be used for categorization as well as loading a data matrix comprising class and their classifications. The input matrix is composed of two classes of elements at once and sent to the algorithm. The subjects from the two classes are displayed in the rows of the matrix, while the features are displayed in the columns. The class is indicated in the final column.Cross-validation, which

divides the data-set into five folds and assesses accuracy, was employed to prevent overfitting mistakes. The Classification Learner gives users the option to select from a variety of classification models, including Logistic Regression: logistic regression,as previously mentioned, is a supervised classification algorithm, which means that it requires as input elements described by a certain feature number and class. The relative accuracy, confusion matrix, scatter plot, and ROC curve of the model are given after model training.

Chapter 3

Results and Discussion

In this chapter, the results obtained from the studies of AVQI and WS are presented and discussed. Besides, it will be discussed whether the GIRBAS perceptual evaluation performed by the experts and the voice parameters produced from the algorithm are consistent. Then, an additional data-set of true healthy subjects (THS) is used to show that the HS subjects provided by speech therapists may not be healthy at the level of the phonatory system but surely they have not MS. Finally, the results of classification algorithm with logistic regression will be shown according to AVQI (VOXplot and Praat), WS and for both by also comparing the three data-set (HS, MS and THS).

3.1 Comparison of software applications to obtain AVQI

Acoustic Voice Quality Index, as it can be seen in paragraph 2.6.1, depends on the parameters: jitter, local shimmer, CPPS, HNR, spectral slope and tilt extracted by a concatenation of 3 seconds of sustained vowel /a/ and 3 seconds of reading task. Four of these parameters, jitter, local shimmer, CPPS, HNR have already been described in the previous chapter. Instead, the *spectral slope* is a basic approximation of the spectrum shape by a linear regression line [29], it refers to the change in signal power or amplitude as the frequency changes. It represents

the decrease of the spectral amplitudes from low to high frequencies. In other words, the spectral slope measures how the amplitude of frequencies changes as it moves from low to high frequencies. A positive spectral slope indicates that the high-frequency components of the spectrum are more intense than the low-frequency components, while a negative spectral slope indicates the opposite. The spectral slope can be calculated using methods such as linear regression on the logarithmic representation of the modulus of the signal spectrum. Furthermore, *Spectral tilt* is a related concept to spectral slope, but often refers specifically to the distribution of energy in different frequency bands. In fact, it is used to describe the overall slope of the spectral power density. In many cases, e.g. in audio signals, higher frequencies have lower power than lower frequencies (1/f characteristic), resulting in a spectral slope. A positive spectral tilt indicates that the lower frequencies are more intense than the higher ones, while a negative spectral tilt indicates the opposite [30]. The extraction of the parameters required to evaluate the index AVQI was done with three software applications: Matlab, Praat and VOXplot.

The spectral slope is a basic approximation of the spectrum shape by a linear regression line. It represents the decrease of the spectral amplitudes from low to high frequencies

1. **Praat:** three seconds of recording of the vowel /a/ renaming it 'sv' as sustained vocal and three seconds of recording of the reading task renaming it 'cs' as continous speech were given as input to the software. Thereafter, a script was used [31], which can be found in Appendix A, in which first the parameters were extracted and then the AVQI was calculated. Tables 3.1 and 3.2 show the extracted parameters for MS and HS, respectively. Instead, the figure 3.1 shows the comparison between AVQI values obtained for the two data sets: for HS, the average value is 5.3 with a standard deviation of 1.4, whereas for MS, the average value is 6.2 with a standard deviation of 1.4.

Table 3.1: Praat results from MS data-set

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)	Slope (dB)	Tilt (dB)	AVQI
CF70	7,9	6,6	20,2	1,7	-24,9	-8,9	7,2
CP46	/	/	/	/	/	/	/
CR60	8,2	9,5	9,5	0,8	-18,6	-7,7	6,2
CS71	10,4	10,7	12,3	1,2	-20,5	-7,9	5,8
DS76	4,5	9,8	19,8	1,5	-32,6	-4,6	9,3
DV62	/	/	/	/	/	/	/
FR51	10,8	12,7	11,1	1,1	-21,0	-8,5	5,1
FS94	/	/	/	/	/	/	/
GF77	12,8	13,0	7,8	0,7	-18,3	-8,3	3,7
GG72	8,9	10,6	15,8	1,4	-26,2	-8,1	6,5
LA71	9,4	8,3	15,6	1,4	-18,91	-9,51	6,1
MC84	7,6	7,5	16,3	1,5	-23,1	-7,4	7,7
MG77	7,9	9,4	15,8	1,5	-26,3	-9,2	6,9
NP69	10,7	11,5	12,9	1,2	-24,4	-8.0	5,3
PM43	8,7	8,8	14,6	1,3	-27,5	-8,2	6,3
SA49	10,5	13,9	9,8	0,9	-25,8	-8,9	4,7

Table 3.2: Praat results from HS data-set

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)	Slope (dB)	Tilt (dB)	AVQI
AM66	11,5	18,4	5,4	0,7	-22,8	-7,5	4,3
BC85	12,3	13,5	6,6	0,7	-14,1	-7,1	4,4
CA68	9,8	10,6	13,8	1,3	-19,7	-9,3	5,7
CC85	8,3	10,8	15,0	1,4	-25,2	-8,7	6,6
CS45	13,6	15,0	9,2	0,8	-20,1	-7,6	3,4
DM51	11,1	13,0	10,2	0,9	-22,8	-7,1	5,0
FG93	/	/	/	/	/	/	/
LU77	9,7	10,9	13,4	1,2	-19,1	-19,1	5,5
MI67	10,1	12,3	10,5	0,9	-25,4	-25,4	5,0
MM73	12,0	14,0	9,6	0,9	-21,4	-21,4	3,9
PM72	8,2	9,2	15,8	1,4	-24,1	-24,1	6,7
PN64	6,1	7,9	17,1	1,42	-26,3	-26,3	8,2
PO65	8,5	10,0	11,7	1,17	-24,9	-24,9	7,0
SA75	/	/	/	/	/	/	/
SM85	12,1	12,5	9,5	0,9	-22,2	-22,2	3,6
VC77	/	/	/	/	/	/	/

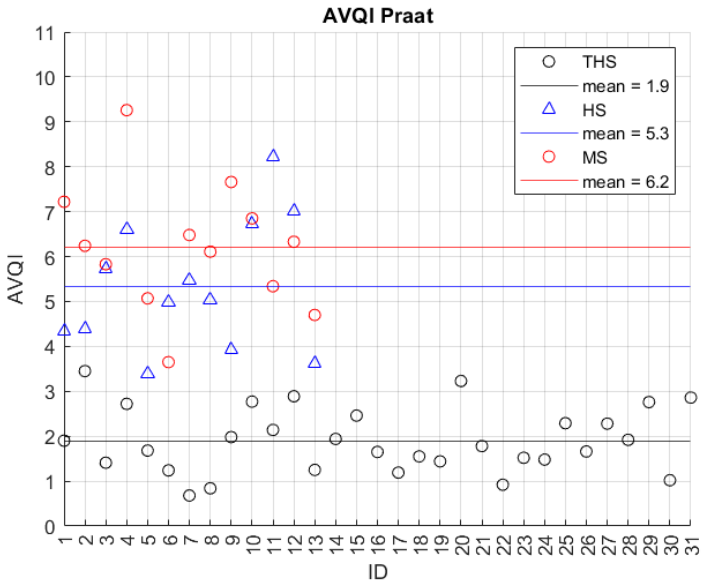

Figure 3.1: Comparison between AVQI value for both HS and MS data sets obtained with Praat

2. **VOXplot:** three seconds of recordings of the sustained vowel /a/ and the reading task were entered directly, i.e. the same recordings were used for all 3 applications (Matlab, VOXplot and Praat). In particular, the recording of sustained vowel is renamed as sustained vowel 'sv'and the reading of reading rask as continous speech 'cs'. From these recordings, the parameters required to calculate the AVQI are returned as output. Tables 3.3 and 3.4 show the extracted parameters for the MS and HS data-sets, respectively.

   The figure 3.2 shows the comparison between AVQI values obtained for HS and MS subjects: for HS, the average value is 3.9 with a standard deviation of 1.5, whereas for MS, the average value is 5.4 with a standard deviation of 2.2.

Table 3.3: VOXplot results from MS data

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)	Slope (dB)	Tilt (dB)	AVQI
CF70	13,8	9,1	17,1	1,5	-23,1	-8,7	5,3
CP46	/	/	/	/	/	/	/
CR60	13,0	11,3	7,9	0,7	-18,3	-6,9	5,1
CS71	16,9	17,4	6,2	0,5	-15,3	-5,5	3,8
DS76	3,5	8,8	16,9	1,5	-29,0	-3,4	10,6
DV62	/	/	/	/	/	/	/
FR51	14,4	16,8	4,6	0,4	-15,9	-9,6	2,9
FS94	/	/	/	/	/	/	/
GF77	14,2	15,3	6,7	0,6	-19,1	-6,6	2,8
GG72	9,5	10,3	21,2	1,8	-34,0	-8,3	6,1
LA71	17,5	12,3	7,9	0,7	-14,5	-8,6	3,2
MC84	6,3	3,6	17,6	1,6	-18,1	-4,9	8,9
MG77	9,0	9,8	23,8	2,0	-38,6	-6,3	7,0
NP69	14,1	12,6	10,5	1,0	-17,5	-6,9	4,6
PM43	11,2	8,7	9,4	0,9	-22,5	-6,3	5,7
SA49	13,1	7,3	14,4	1,3	-14,6	-4,6	5,0

Table 3.4: VOXplot results from HS data

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)	Slope (dB)	Tilt (dB)	AVQI
AM66	12,3	23,0	3,9	0,4	-23,2	-8,3	4,0
BC85	14,3	15,5	4,4	0,4	-12,0	-7,0	3,3
CA68	16,0	17,9	5,8	0,5	-16,4	-8,5	3,1
CC85	14,3	12,2	11,3	1,0	-13,1	-9,5	4,3
CS45	16,4	21,8	2,5	0,2	-22,2	-7,6	1,2
DM51	13,23	16,41	4,45	0,42	-18,5	-8,6	3,4
FG93	/	/	/	/	/	/	/
LU77	14,1	12,2	17,0	1,5	-15,3	-13,1	3,9
MI67	13,6	15,4	4,1	0,4	-18,1	-7,8	3,2
MM73	10,9	16,2	7,5	0,7	-22,4	-7,7	4,3
PM72	10,5	11,0	14,2	1,3	-23,2	-7,3	5,7
PN64	9,7	10,6	15,8	1,4	-21,7	-4,6	7,3
PO65	12,8	9,8	11,6	1,1	-22,6	-4,8	5
SA75	/	/	/	/	/	/	/
SM85	16,0	14,5	6,9	0,6	-19,4	-7,4	2,0
VC77	/	/	/	/	/	/	/

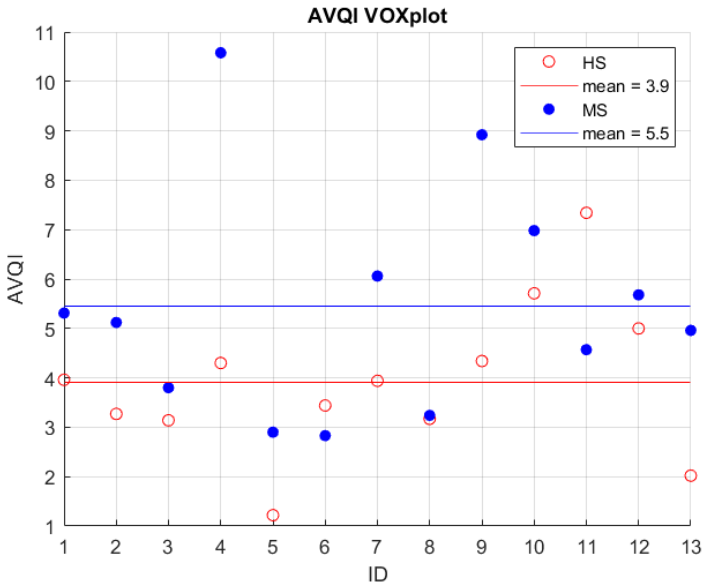

Figure 3.2: Comparison between AVQI values obtained for HS and MS data-sets obtained with VOXplot

3. **MATLAB:** concerning the results of this application,the comparison is also to be made with AVQI obtained from MATLAB, and for this reason a script was developed that extracts from the file 3 seconds from the sustained speech recording and 3 seconds from the continous speech recording the parameters that contribute to the evaluation of AVQI. Tables 3.5 and 3.6 show the extracted parameters respectively for MS and HS but not the AVQI, while Figure 3.3 and 3.4 show a comparison of the various parameters in these tables between the 3 software applications (Praat, VOXplot and Matlab) respectively for MS and HS data-set.

Table 3.5: Matlab results from MS data-set

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)
CF70	12,7	8,6	8,6	0,8
CP46	/	/	/	/
CR60	12,7	11,3	10,6	1,0
CS71	15,4	17,2	3,7	0,4
DS76	7,1	7,9	13,4	1,2
DV62	/	/	/	/
FR51	13,4	13,2	7,3	0,7
FS94	/	/	/	/
GF77	15,5	14,9	5,3	0,5
GG72	12,3	7,6	15,3	1,5
LA71	13,7	11,5	9,3	0,9
MC84	11,1	8,4	15,2	1,7
MG77	9,5	7,0	27,9	3,4
NP69	12,0	11,2	9,9	1,1
PM43	9,3	9,7	12,8	1,2
SA49	12,0	10,5	14,4	1,4

Table 3.6: Matlab results from HS data-set

ID	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimmer (%)	Shimmer (dB)
AM66	14,3	18,0	4,4	0,4
BC85	14,3	15,1	5,1	0,5
CA68	14,6	13,7	7,4	0,7
CC85	13,8	12,2	7,7	0,7
CS45	15,6	14,2	8,2	0,7
DM51	12,7	11,7	9,4	0,8
FG93	/	/	/	/
LU77	12,0	12,5	13,2	1,2
MI67	14,3	14,1	7,0	0,6
MM73	13,5	13,8	7,9	0,7
PM72	13,0	10,4	12,1	1,0
PN64	13,6	11,5	9,3	0,9
PO65	12,7	10,5	10,4	0,9
SA75	/	/	/	/
SM85	15,1	13,3	8,2	0,7
VC77	/	/	/	/

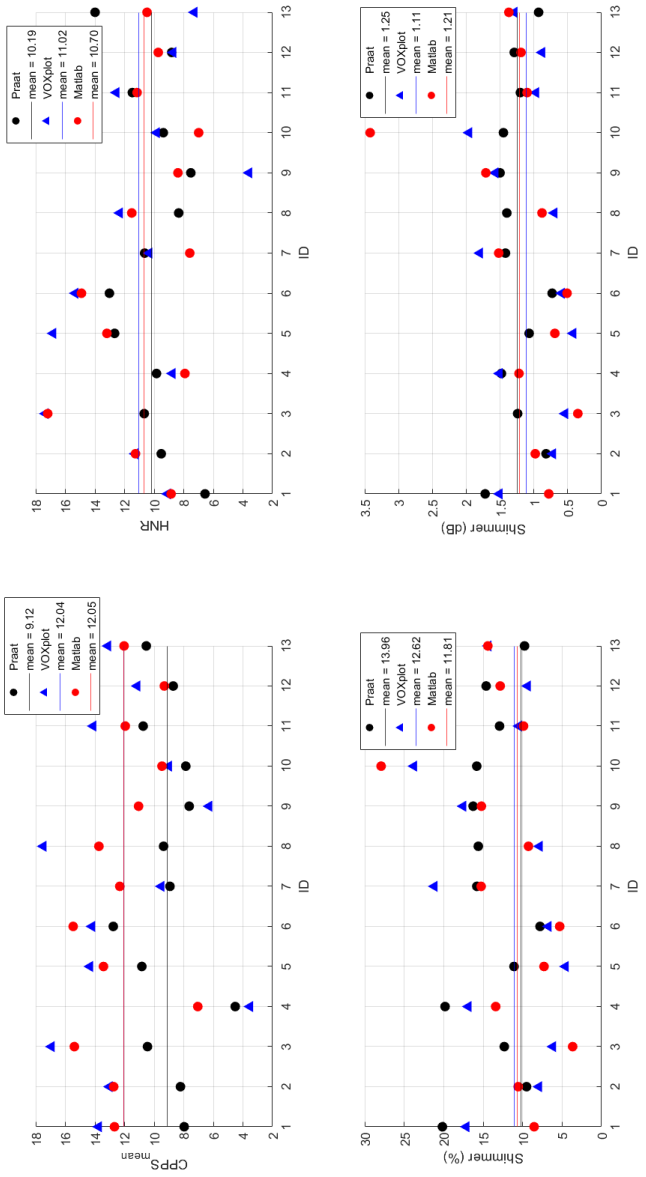

Figure 3.3: AVQI Parameters of MS

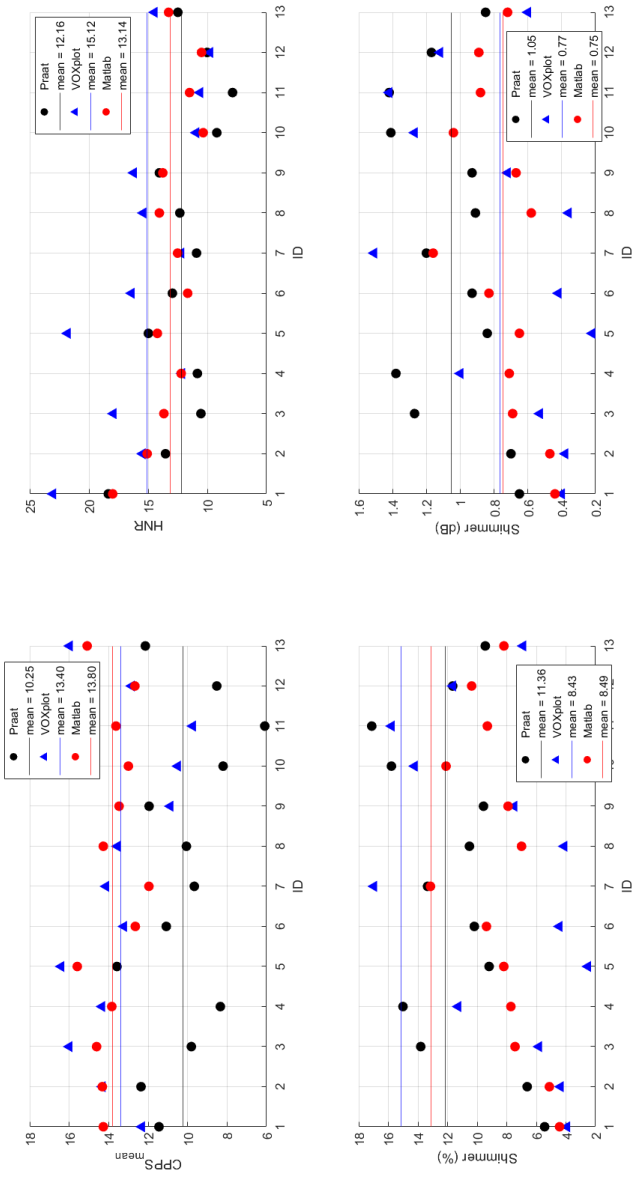

Figure 3.4: AVQI Parameters of HS

3.2 Warning Score

A novel index called Warning Score was used to evaluate each subject's vocal health state. From the 3 repetitions of the sustained vowel /a/, the average values of the parameters, local jitter (%), local shimmer (%), and mean and standard deviation of the Cepstral Peak Prominence smoothed CPPS (dB), are taken into consideration. It was assigned based on parameters taken from both the VH and the MI recordings. Depending on whether the subject was pathological or healthy, positive or negative scores were assigned after examining the recovered parameters. In particular, a positive value is assigned for pathological voice and negative value for a healthy one.
As mentioned in paragraph 2.6.2, Warning Score is given according to the following rule:

- A value less than 0.31% suggests a healthy voice, whereas a number greater than 0.43% suggests a pathological voice. If the local jitter value is between (0.31 - 0.43)%, it is considered unreliable.
- A local shimmer value of less than 2.37% indicates a healthy voice, more than 2.55% indicates a pathological voice, and if it varies between (2.37 - 2.55)%, it is deemed unreliable.
- The $\mathrm{CPPS_{mean}}$ value is considered unreliable if it is between (18.0 - 19.7) dB, a value less than 18.0 dB indicates a pathological voice, and value higher than 19.7 dB indicates a healthy voice.
- A $\mathrm{CPPS_{std}}$ value of less than 0.9 dB indicates a healthy voice, greater than 1.3 dB indicates a pathological voice, and if it falls between (0.9 - 1.3) dB, it is deemed unreliable.

Table 3.7 and 3.8 show the WS results respectively with MI and VH recordings for multiple sclerosis subject. From these results, it is possible to observe concordant results for both VH and MI. In fact, a high WS value indicates a pathological state, and in this case it can be seen that all subjects received a WS between 3 and 4.

Table 3.7: WS results from MS with MI

ID	Jitter (%)	Shimmer (%)	$\mathbf{CPPS_{mean}}$ (dB)	$\mathbf{CPPS_{std}}$ (dB)	Warning Score
CF70	0,9	7,8	15,5	2,1	4
CP46	/	/	/	/	/
CR60	/	/	/	/	/
CS71	0,5	5,0	16,4	1,7	4
DS76	6,1	15,6	6,6	1,9	4
DV62	/	/	/	/	/
FR51	0,8	5,2	14,2	1,8	4
FS94	/	/	/	/	/
GF77	0,4	4,9	16,2	1,5	3
GG72	0,8	7,5	13,9	1,6	4
LA71	0,8	8,2	15,0	1,8	4
MC84	2,7	23,7	9,5	2,4	4
MG77	1,1	39,5	8,2	1,8	4
NP69	0,8	7,3	16,2	1,7	4
PM43	4,2	14,1	14,8	2,5	4
SA49	1,5	10,9	14,1	2,4	4

Table 3.8: WS results from MS with VH

ID	Jitter (%)	Shimmer (%)	$\mathbf{CPPS_{mean}}$ (dB)	$\mathbf{CPPS_{std}}$ (dB)	Warning Score
CF70	3,0	8,6	15,6	2,4	4
CP46	1,5	11,2	16,3	2,1	4
CR60	1,0	16,3	15,8	2,4	4
CS71	2,1	3,7	18,0	3,9	3
DS76	2,2	7,8	11,3	2,0	4
DV62	1,7	8,1	8,9	1,1	3
FR51	2,1	5,1	16,9	3,5	4
FS94	1,0	4,9	12,1	1,5	4
GF77	/	/	/	/	/
GG72	0,4	2,8	14,9	1,6	4
LA71	0,8	4,7	16,8	1,7	4
MC84	6,4	10,8	11,6	3,1	4
MG77	1,1	3,3	12,5	1,8	4
NP69	0,4	2,9	16,7	1,6	3
PM43	8,5	25,4	17,5	3,6	4
SA49	1,3	4,1	15,7	2,5	4

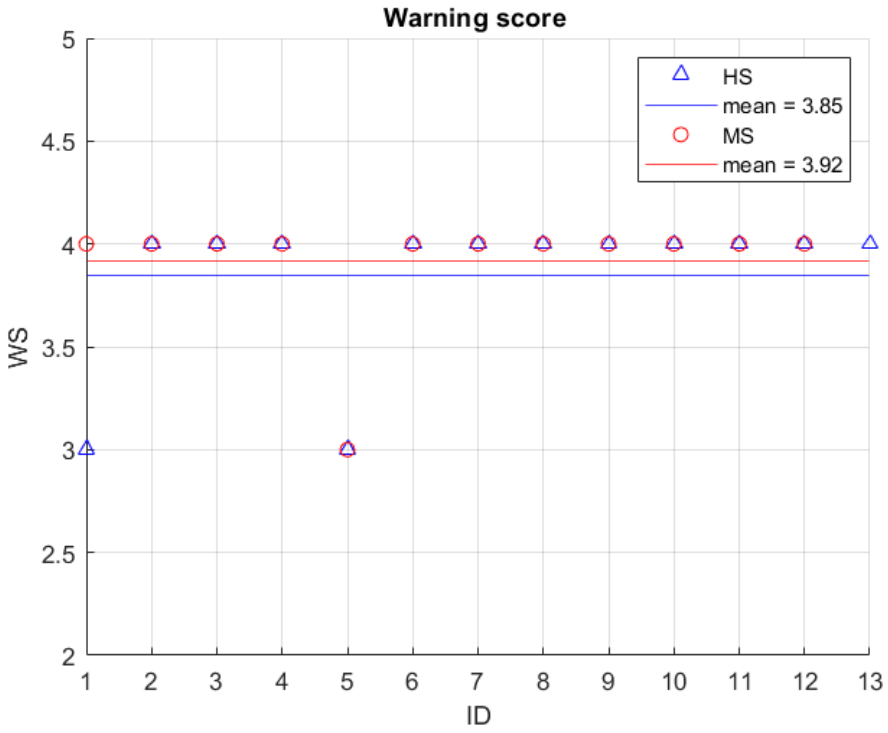

Figure 3.5: Comparison MI recording between HS and MS of Warning Score

Table 3.9 and 3.10 show the WS results respectively with MI and VH recordings for healthy subject. Differently from the MS subjects, for the HS there is a difference in the results between the VH and MI recordings. In fact, in the MI recordings (table 3.10), a WS between 3 and 4 were given to subjects, whereas for the VH recordings (table 3.9), only one subject had a WS of -1, which indicates actual health status, but most of the HS had a WS between 3 and 4, it can also be seen in Figure 3.5. For this reason, it started to be assumed that HS subjects were not really healthy but might have other diseases of the phonatory system than MS. Therefore, another data-set of true healthy subjects (THS) was later added, which was used as a control group compared to MS, also for classification.

Table 3.9: WS results from HS with VH

ID	Jitter (%)	Shimmer (%)	$\mathbf{CPPS_{mean}}$ (dB)	$\mathbf{CPPS_{std}}$ (dB)	Warning Score
AM66	0,21	1,2	15,8	1,0	-1
BC85	/	/	/	/	/
CA68	0,67	1,7	19,3	1,8	1
CC85	1,08	5,8	13,2	1,3	4
CS45	0,70	2,0	18,1	2,2	1
DM51	2,18	8,1	16,4	3,3	4
FG93	/	/	/	/	/
LU77	0,67	8,5	14,7	2,7	4
MI67	0,39	3,1	17,2	1,6	3
MM73	0,88	3,3	16,0	1,8	4
PM72	0,45	3,2	17,1	1,9	4
PN64	0,45	3,9	15,6	1,8	4
PO65	0,69	5,6	16,2	1,9	4
SA75	0,35	3,6	15,3	1,3	3
SM85	0,40	6,3	15,9	1,1	2
VC77	0,34	1,7	18,6	1,4	0

Table 3.10: WS results from HS with MI

ID	Jitter (%)	Shimmer (%)	CPPS (dB)	CPPS (dB)	Warning Score
AM66	0,32	3,0	16,0	1,5	3
BC85	0,43	4,3	15,7	1,4	4
CA68	0,51	10,6	17,3	1,7	4
CC85	0,54	7,9	16,5	1,8	4
CS45	3,01	7,1	18,5	2,3	3
DM51	1,54	8,1	15,5	2,8	4
FG93	/	/	/	/	/
LU77	1,01	12,4	14,5	2,3	4
MI67	0,96	8,3	14,7	2,1	4
MM73	0,91	5,4	16,6	1,9	4
PM72	0,68	10,3	14,7	2,1	4
PN64	0,65	9,8	13,4	2,1	4
PO65	0,80	14,5	15,2	2,4	4
SA75	/	/	/	/	/
SM85	1,70	10,3	16,2	1,7	4
VC77	/	/	/	/	/

3.2.1 Comparison of MI and VH recordings between HS and MS

Acquisitions of the sustained vowel /a/ were made with both the air microphone (MI) and the contact microphone (VH). In a previous thesis [32], an attempt was made to find a correlation between these two types of recordings and their extracted parameters. The analysis is carried out by computing the differences, denoted by the Greek capital letter Δ, between the parameters that were extracted from the air microphone (referred to in this study as MIC for short) and the parameters that were obtained with VH and stored inside the DAP unit. Deltas give information about the degree to which the parameters processed by the contact microphone-based device and those derived from the in-air microphone differ. The metrics that the in-air microphone and VH share, when taking into account the vowel /a/ task repeats, are local jitter (%), local shimmer (%), $\text{CPPS}_{\text{mean}}$ (dB), and CPPS_{std} (dB). Although the variances between the two microphones are not insignificant, it is not expected for delta values to equal zero because each subject's two input signals differ. The VH uses the mechanical signal produced by the vocal folds at the neck, which are thought of as a low-pass filter, as its input signal. In contrast, the in-air microphone records an in-air pressure signal that is modulated by the vocal tract. Furthermore, the two devices use distinct measurement chains. The usage of the VH device needs the definition of precise cut-off values for the extracted parameters due to differing features of the devices. The VH device is used because it is more convenient to use during acquisitions, allowing the subject to move freely without worrying about the distance between their lips and the microphone, and because it is less sensitive to other potential sound sources in the surrounding area.

3.3 Comparison between HS and MS subjects and GIRBAS scale

The association between the GIRBAS scale determined by the experts and the outcome of HS and MS subjects is the subject of another analysis carried out in this study. Several techniques for the auditory-perceptual judgments are available to assess voice quality, including the GIRBAS scale. In particular, only the G

and A parameters of the GIRBAS scale were taken into account because these are the studies that showed the most evidence. The variations in ratings among different listeners or even within a single listener serve as a primary indicator of voice quality [33]. Figures 3.6 and 3.7 show no correspondence with the GIRBAS scale and WS values for HS and MS data-set and this result confirms previous studies [34]. In fact, most subjects had a GIRBAS value of 0, which corresponds to an optimal state of health. However, there is no substantial difference between the GIRBAS values assigned to HS and MS. In the following figures, the x-axis shows the value of G and A on the GIRBAS scale assigned by the speech therapists, while the y-scale shows the value of the respective parameters. In particular, the yellow line identifies the limit beyond which the subject is considered unhealthy and is assigned a positive WS, while the green line identifies the limit value beyond which the subject is considered healthy and is assigned a negative WS.

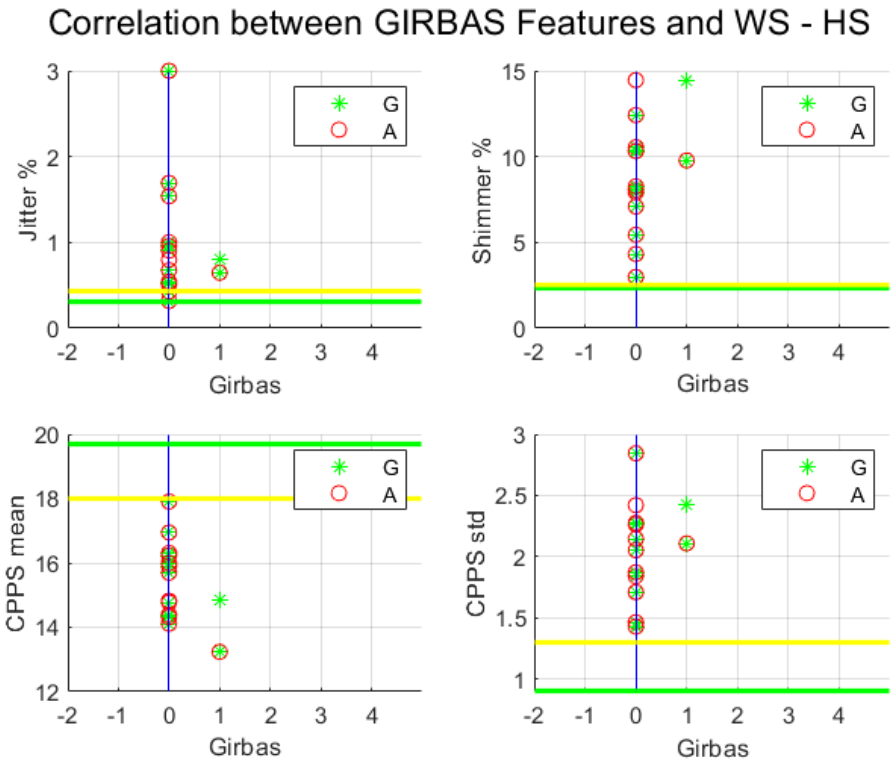

Figure 3.6: Correlation between WS parameters and GIRBAS for HS

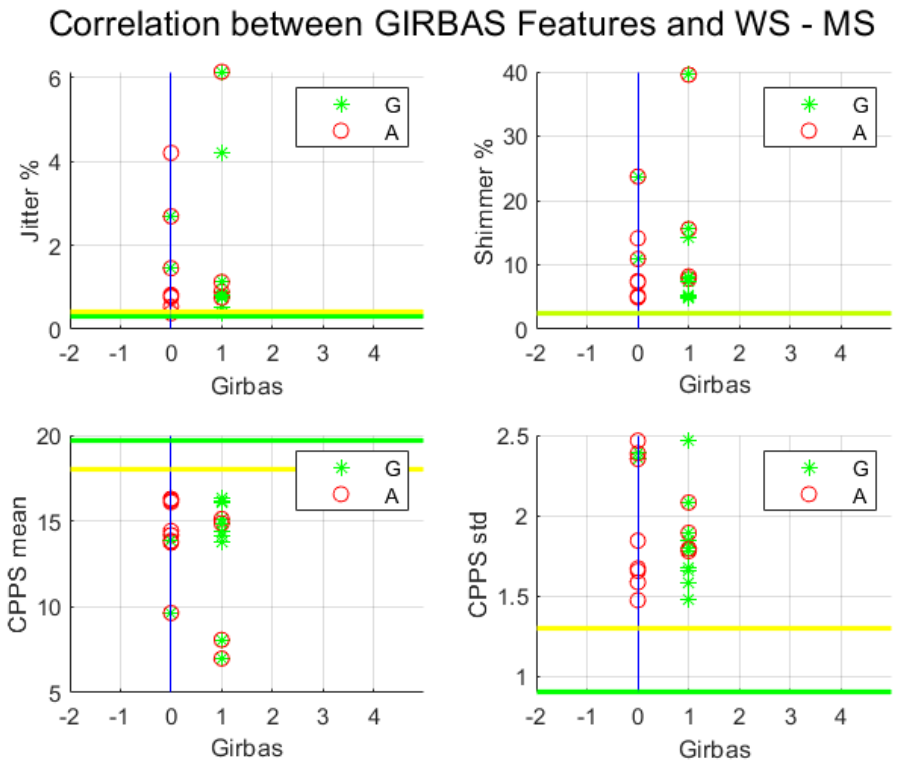

Figure 3.7: Correlation between WS parameters and GIRBAS for MS

3.4 Classification

After extracting the parameters and evaluating the subjects according to their AVQI and WS index, the subjects were classified according to the two index evaluating the performance. Classification, for convenience, was carried out entirely within the 'Classification Learner' application. When loading the data matrix containing all the available data, the setting was chosen to perform a validation phase by means of a 5 fold cross-validation. This allowed the data-set to be divided into five equivalent portions in terms of capacity and in each one, training was performed on some observations and validation on others totally different from the former. Finally, the logistic regression model was selected in the relevant menu, and the index feature (AVQI or WS) was selected, which reported the best classification results, as shown in Table 3.11, 3.12 and 3.13. The two classes included in this work are displayed in the upper left corner of the tables; the positive class is tied to MS

patients, and the negative class is associated with healthy participants. In these tables Area Under the Curve (AUC), Precision, Sensitivity, Specificity, F1-score and Accuracy are shown. In terms of AVQI, only parameters extracted with Praat and VOXplot were used: subjects were classified using the Logistic Regression (LR) model described in paragraph 2.8. Metrics related to the classification based on the AVQI index calculated with Praat are displayed in Table 3.11, and metrics related to the classification based on the AVQI index computed with VOXplot are displayed in Table 3.12. It is evident that the results acquired with VOXplot outperform the ones obtained using Praat. The performances that were obtained based on the classification according to Warning Score (table 3.13) are significantly worse. This shows that the HS are not healthy subjects, but rather that they do not have multiple sclerosis; in fact, their bad outcomes could be caused by various phonatory system disorders. In addition, the characteristics of the model of Praat AVQI were shown by means of the confusion matrix, in figure 3.8, and the ROC curve in figure 3.9.

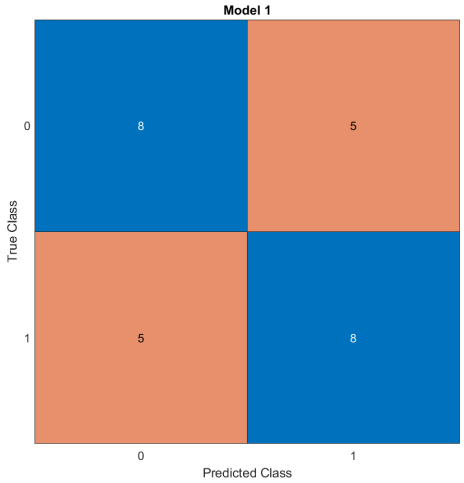

Figure 3.8: Confusion matrix of logistic regression model of Praat AVQI

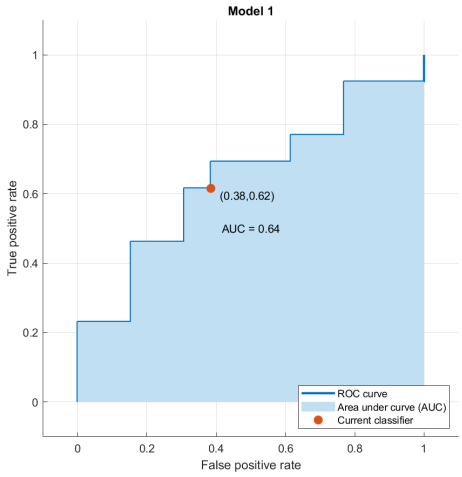

Figure 3.9: Area under ROC curve of logistic regression model of Praat AVQI

Table 3.11: Classification performance obtained for AVQI by Praat application

HS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
AVQI	54,0%	61,5%	61,5%	61,5%	61,5%	61,5%

The ROC curve in figure 3.11 and the confusion matrix in figure 3.10 were used to illustrate the features of the VOXplot AVQI model. High sensitivity value means a correct classification for a large proportion of class 0 (HS) subjects, and also high AUC value expresses a high quality of classification of the predictions obtained. In this case, high value of sensitivity and AUC were obtained, so VOXplot can be defined as the best application to calculate AVQI but it could be better (paragraph 3.5).

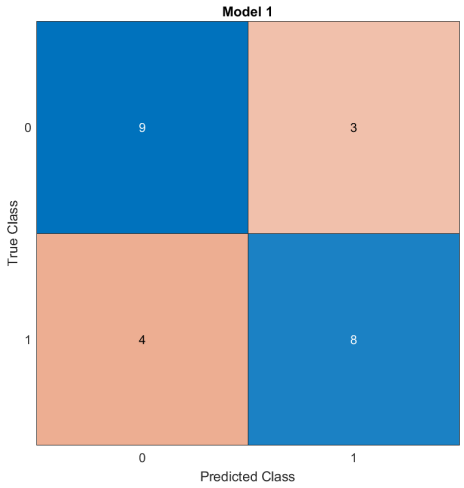

Figure 3.10: Confusion matrix of logistic regression model of VOXplot AVQI

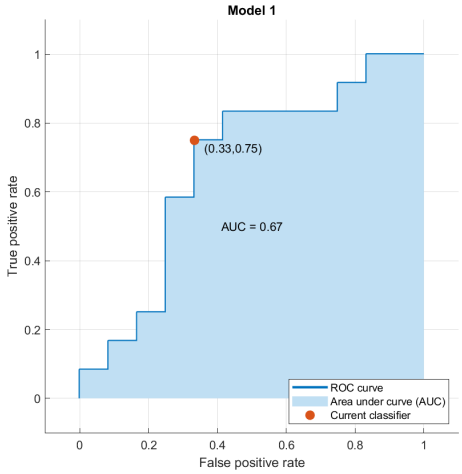

Figure 3.11: Area under ROC curve of logistic regression model of VOXplot AVQI

Table 3.12: Classification performance obtained for AVQI by VOXplot application

HS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
AVQI	63,0%	69,2%	75,0%	66,7%	72,0%	70,8%

Furthermore, figures 3.12 and 3.13 depict the confusion matrix and ROC curve, respectively, which illustrate the features of the WS model. The latter creates a curve that illustrates the model performance of the model for any classification threshold that is selected by plotting the sensitivity of classifier value against its specificity. This makes it easy to visually recognize the features of the current model, including the elements that are classified as positive corrected and negative corrected, as well as the various features that are achievable with various classification levels.

Additionally, the AUC (Area Under Curve), which calculates the area occupied by the ROC curve and evaluates the ability of the model to differentiate between the two classes under investigation, can be highlighted in such a ROC curve. Furthermore, Table 3.13 shows the accuracy value of 41.7% and especially the AUC of 36%, which imply low classification performance between HS and MS in the case of classifying subjects according to the Warning Score.

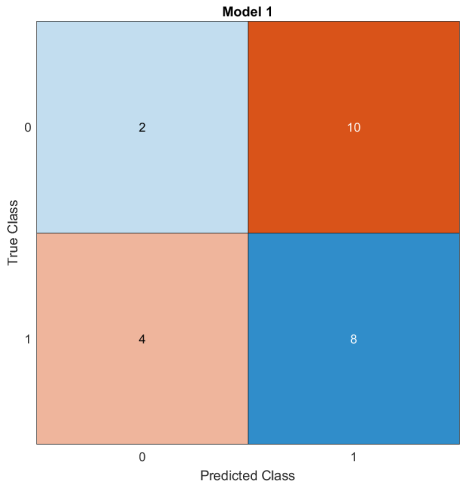

Figure 3.12: Confusion matrix of logistic regression model of Warning Score between HS and MS

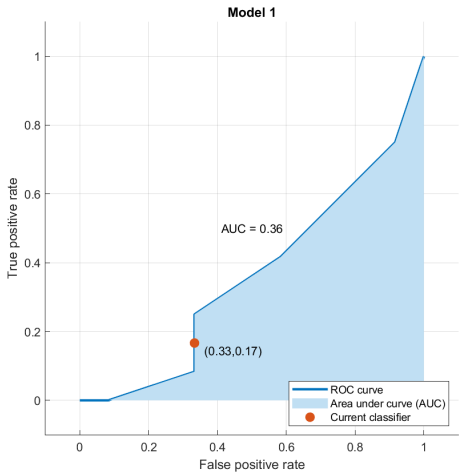

Figure 3.13: Area under ROC curve of logistic regression model of Warning Score between HS and MS

Table 3.13: Classification performance obtained for WS

HS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
Warning Score	36,0%	16,7%	33,3%	66,7%	22,2%	41,7%

3.5 True Healthy Subjects Data-set

From the results obtained previously, it was thought that the HS were not really healthy subjects but simply not free of other phonatory pathologies. Therefore, it was decided to include an additional data-set with totally healthy phonatory subjects as a control group called True Healthy Subjects (THS). This data-set was provided by the Department of Electronics and Telecommunications of the Polytechnic University of Turin and includes 57 subjects, 29 males and 28 females. For these subjects, 3 repetitions of the sustained vowel /a/, reading and free speech task were provided.
Similar to the previous data-sets, the same analyses were carried out: first the WS was assigned, then the AVQI was found with both Praat and VOXplot and finally the subjects were classified.

Table 3.14: WS results from THS data

ID	GENDER	Jitter (%)	Shimmer (%)	$CPPS_{mean}$ (dB)	$CPPS_{std}$ (dB)	Warning Score
1	M	0,38	4,0	17,1	1,6	3
2	M	0,30	2,7	18,2	1,2	0
3	M	0,21	2,7	18,2	1,3	0
4	M	0,23	1,9	17,6	1,4	0
5	F	0,20	1,1	14,5	1,1	-1
6	F	0,28	1,9	17,0	1,3	0
7	F	0,13	1,0	14,4	1,2	-1
8	F	0,21	1,3	15,6	1,5	0
9	M	0,40	2,2	18,2	1,1	-1
10	M	0,33	2,2	16,8	1,3	0
12	F	0,23	1,7	14,9	1,3	-1
13	F	0,34	3,1	16,5	1,3	3
15	F	0,24	1,6	18,8	1,3	-2
17	F	0,21	1,5	18,7	1,5	-1
18	M	0,34	1,7	17,7	1,1	0
19	M	0,26	2,0	15,9	1,2	-1
20	F	0,21	1,3	12,8	0,9	-2
21	M	0,26	2,8	18,8	1,1	0
22	M	0,22	1,4	18,3	1,5	-1
23	M	0,15	1,4	18,3	1,5	-1
24	M	0,44	1,8	18,0	1,2	0
25	M	0,39	2,8	18,5	1,2	1

ID	**GENDER**	**Jitter (%)**	**Shimmer (%)**	**$CPPS_{mean}$ (dB)**	**$CPPS_{std}$ (dB)**	**Warning Score**
26	M	0,18	1,8	16,3	1,2	-1
27	F	0,37	2,3	14,3	1,2	0
28	F	0,14	1,3	15,2	1,4	0
29	F	0,21	1,3	19,9	1,4	-2
30	M	0,41	2,0	17,9	1,1	0
31	M	0,33	2,8	18,9	1,2	1
33	M	0,34	3,4	18,4	1,3	2
34	M	0,29	2,8	17,3	1,1	1
35	F	0,38	3,1	17,3	1,4	3
36	M	0,37	3,0	15,2	1,2	2
37	F	0,41	4,0	17,8	1,5	3
38	M	0,34	3,9	19,3	1,2	1
39	M	0,38	2,7	17,3	1,3	3
40	F	0,23	3,0	14,6	1,3	1
41	F	0,24	1,4	17,6	1,1	-1
46	M	0,34	4,1	16,2	1,5	3
48	F	0,16	2,3	15,2	1,2	-1
50	F	0,27	2,9	16,2	1,2	1
51	F	0,57	3,0	17,7	1,5	4
56	M	0,30	2,4	15,7	1,4	0
60	F	1,22	6,5	15,2	2,2	4
61	F	0,32	2,2	15,8	1,5	1
76	F	0,29	2,2	14,6	1,7	0
81	F	0,74	4,4	15,2	1,3	4
82	F	0,66	5,6	15,0	1,5	4
84	F	1,14	6,4	14,8	1,6	4
85	F	0,51	3,4	15,1	1,2	3
97	F	0,33	2,6	18,5	1,1	1
100	M	0,33	3,9	16,2	1,0	2
101	F	0,35	3,2	18,4	1,3	1
102	M	0,10	1,1	5,6	0,5	-3
103	M	0,83	10,6	17,3	1,9	4
104	M	0,32	2,9	18,2	1,4	2
105	M	0,19	1,6	19,0	1,6	-1
106	M	0,31	1,2	19,0	1,3	-3

Similar to the previous data-set, the parameters $CPPS_{mean}$, $CPPS_{std}$, jitter and shimmer were extracted and the Warning Score was assigned according to the criterion seen in section 2.7. As can be seen in table 3.14, 27 subjects had a WS greater than 0 resulting 'pathologic', 17 had a WS less than 0 resulting 'healthy'

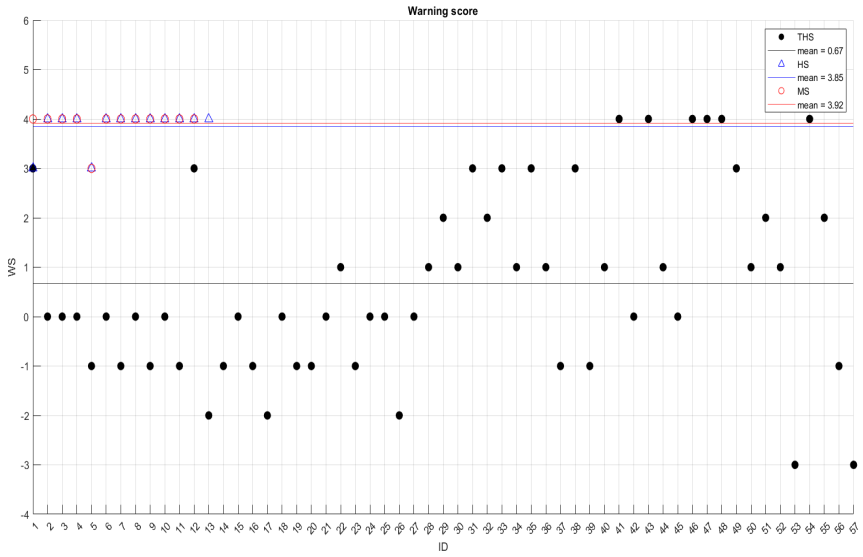

Figure 3.14: Comparison among THS, HS and MS data-set through the parameter Warning Score

and 13 had a WS equal to 0 resulting in an 'indeterminate'.
Figure 3.14 displays the WS values for each of the HS, MS and THS subjects. In particular, for the HS the average value is 3.85, for the MS is 3.92, and finally for the THS the average value is 0.67, thus demonstrating that THS are really in the 'healthy' range and can be used as a control group.

Furthemore, Figure 3.15 shows the relative occurrences of the Warning Score for the three categories (HS, MS and THS). The relative occurrence is defined as the ratio between the number of times the WS is equal to a value and the number of all subjects: it can be seen that for HS and MS all subjects had a WS between 3 and 4. The same analysis for the THS provided the results shown in Figure 3.15, where it can be seen that 29.3% of the subjects had a WS between -4 and -1 resulting in the case of 'healthy', 24.2% had a value of 0 resulting in the case of 'unreliable' and the remaining 46.5% had a result between 1 and 4 resulting in the case of 'pathological'.

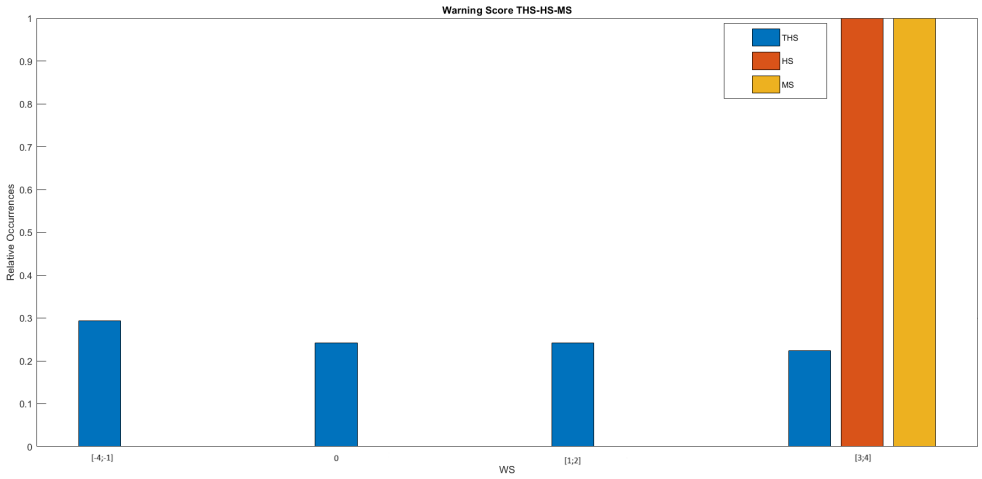

Figure 3.15: Relative occurrences of THS, HS and MS's Warning Score

For the calculation of the AVQI with the two software applications, only THS who received a WS between -4 and 0 were considered (31 subjects).

Table 3.15: AVQI results of THS data-set from Praat

ID	GENDER	$\mathbf{CPPS_{mean}}$ (dB)	HNR (dB)	Shimm (%)	Shimm (dB)	Slope (dB)	Tilt (dB)	AVQI
2	M	14,1	14,3	7,4	0,8	-20,7	-12,3	1,9
3	M	11,5	14,1	8,9	0,9	-22,0	-12,1	3,5
4	M	14,9	15,8	6,7	0,8	-18,1	-13,0	1,4
5	F	13,4	16,5	5,3	0,8	-13,4	-12,8	2,7
6	F	13,9	16,8	5,2	0,6	-17,7	-12,9	1,7
7	F	15,1	15,7	5,3	0,7	-16,5	-12,9	1,2
8	F	16,7	16,7	6,0	0,7	-14,7	-12,2	0,7
9	M	15,6	15,0	6,6	0,7	-19,1	-13,3	0,8
10	M	13,9	16,1	6,8	0,7	-19,3	-12,1	2,0
12	F	12,6	16,9	7,6	0,9	-20,5	-12,7	2,8
15	F	13,6	16,0	5,6	0,6	-20,0	-11,8	2,1
17	F	12,3	13,3	9,02	0,9	-21,3	-12,5	2,9
18	M	15,6	15,2	6,4	0,7	-18,0	-12,2	1,3
19	M	14,4	17,2	6,2	0,7	-18,1	-11,8	1,9
20	F	12,9	13,7	7,3	0,8	-20,4	-12,7	2,5
21	M	14,2	16,8	7,0	0,7	-21,9	-12,4	1,7
22	M	14,9	15,3	5,9	0,6	-19,2	-12,6	1,2
23	M	14,4	15,9	6,9	0,8	-19,1	-13,3	1,6
24	M	14,7	14,7	5,8	0,6	-19,6	-12,4	1,4
26	M	11,6	16,8	4,5	0,6	-16,9	-12,1	3,2
27	F	14,5	17,5	5,5	0,7	-13,8	-12,7	1,8
28	F	15,9	19,2	4,6	0,6	-15,2	-12,7	0,9
29	F	14,9	15,9	3,6	0,6	-13,2	-12,7	1,5
30	M	14,7	15,3	6,4	0,7	-21,8	-12,0	1,5
41	F	13,6	12,7	8,4	0,9	-17,5	-12,8	2,3
48	F	14,3	17,9	6,6	0,7	-17,0	-12,5	1,7
56	M	13,7	14,8	7,08	0,8	-14,8	-13,0	2,3
76	F	14,7	15,6	8,7	0,8	-14,2	-12,1	1,9
102	M	12,2	12,6	8,9	0,8	-19,8	-13,2	2,8
105	M	16,1	14,2	6,1	0,8	-16,6	-13,0	1,0
106	M	12,7	13,4	4,8	0,6	-17,3	-11,1	2,9

Table 3.16: AVQI results of THS data-set from VOXplot

ID	GENDER	$CPPS_{mean}$ (dB)	HNR (dB)	Shim (%)	Shim (dB)	Slope (dB)	Tilt (dB)	AVQI
2	M	20,4	23,8	3,9	0,3	-17,9	-10,2	0,0
3	M	17,3	25,9	2,4	0,2	-17,9	-13,7	0,6
4	M	19,8	24,9	2,4	0,2	-14,6	-12,7	-0,6
5	F	16,3	28,3	1,4	0,1	-12,2	-10,1	1,5
6	F	19,2	24,1	2,2	0,9	-13,0	-10,9	-0,4
7	F	19,3	33,3	0,7	0,1	-1,6	-9,8	-0,7
8	F	22,7	30,1	1,0	0,1	-14,2	-7,6	-1,6
9	M	20,1	24,9	1,9	0,2	-16,7	-12,4	-1,1
10	M	17,5	25,9	2,0	0,1	-13,8	-10,5	0,2
12	F	18,1	30,1	1,6	0,1	-19,3	-10,6	0,2
15	F	21,0	28,8	0,8	0,1	-15,0	-9,5	-0,8
17	F	20,9	28,1	1,3	0,1	-16,9	-11,4	-0,6
18	M	23,3	30,1	1,3	0,1	-15,7	-11,5	-2,2
19	M	16,8	25,5	1,9	0,2	-10,4	-10,2	0,9
20	F	19,7	23,1	1,9	0,2	-15,9	-11,9	-0,5
21	M	18,4	31,2	1,8	0,2	-22,6	-9,7	-0,2
22	M	21,3	27,3	1,5	0,1	-17,0	-12,5	-1,7
23	M	20,5	31,9	1,4	0,1	-14,8	-11,0	-0,8
24	M	20,9	22,5	1,7	0,2	-16,4	-13,6	-0,8
26	M	17,7	28,4	2,5	0,2	-17,3	-10,3	0,6
27	F	21,1	30,4	1,4	0,1	-10,4	-10,6	-1,2
28	F	18,6	29,4	1,7	0,2	-9,6	-11,7	0,0
29	F	20,1	28,9	1,1	0,1	-11,9	-10,2	-0,7
30	M	19,8	23,5	1,7	0,2	-18,9	-10,6	-0,1
41	F	21,2	27,5	1,4	0,1	-14,6	-11,2	-1,0
48	F	15,9	23,9	2,2	0,2	-13,3	-8,9	0,28
56	M	17,2	26,3	2,1	0,2	-9,6	-12,4	0,6
76	F	18,0	27,1	1,6	0,1	-12,8	-8,0	0,2
102	M	18,6	19,6	3,3	0,3	-14,3	-13,0	0,6
105	M	22,7	27,8	1,2	0,1	-15,3	-11,0	-1,4
106	M	19,9	26,4	0,8	0,1	-14,1	-10,1	-0,1

Tables 3.15 and 3.16 show the extracted parameters for THS respectively from Praat and VOXplot application. Instead, figure 3.16 and 3.17 show the comparison among THS, MS and HS data-set of AVQI for each subjects: for Praat, the average

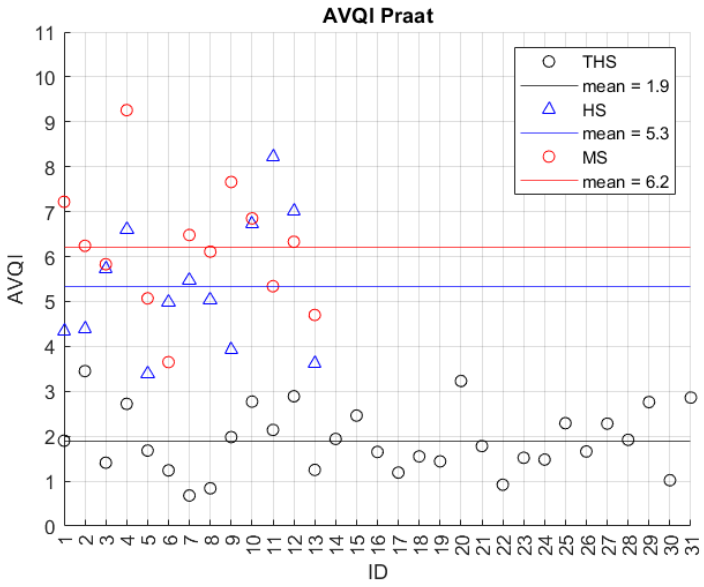

Figure 3.16: Comparison among THS, HS and MS data-set of results of AVQI from Praat

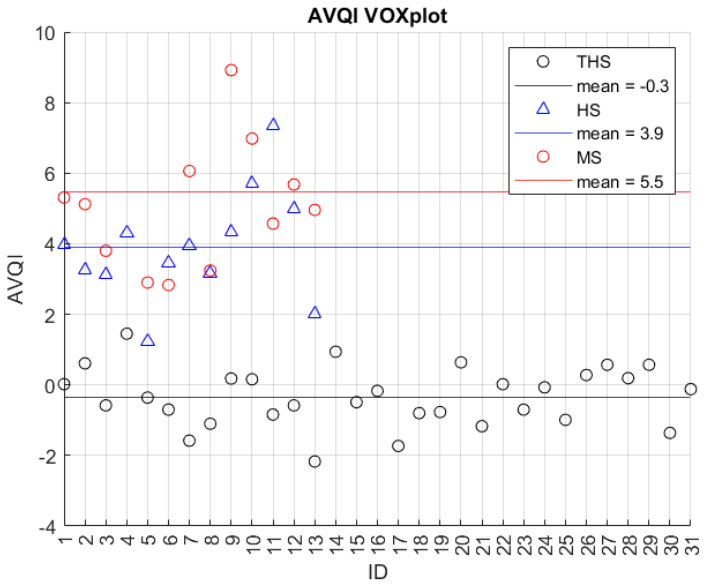

Figure 3.17: Comparison among THS, HS and MS data-set of results of AVQI from VOXplot

value of THS is 1.9 with a standard deviation of 0.70, whereas for VOXplot, the average value of THS is -0.34 with a standard deviation of 0.81.

Table 3.17: Matlab results from THS data-set

ID	**GENDER**	**$CPPS_{mean}$ (dB)**	**HNR (dB)**	**Shimmer (%)**	**Shimmer (dB)**
2	M	17,2	10,4	9,9	0,98
3	M	15,7	15,9	5,3	0,50
4	M	17,1	18,1	2,7	0,26
5	F	14,5	21,3	0,9	0,08
6	F	15,9	17,3	3,1	0,30
7	F	14,4	23,6	1,3	0,12
8	F	15,7	20,0	2,1	0,21
9	M	17,4	12,8	9,4	0,83
10	M	15,9	10,6	12,6	1,44
12	F	15,3	21,4	1,8	0,16
15	F	18,2	22,0	1,4	0,15
17	F	18,1	20,5	2,4	0,28
18	M	17,5	20,0	1,6	0,16
19	M	14,7	17,6	3,1	0,32
20	F	19,3	18,9	2,2	0,19
21	M	18,3	17,9	3,8	0,38
22	M	17,7	19,3	3,0	0,32
23	M	17,5	20,7	1,6	0,16
24	M	16,6	18,0	3,9	0,38
26	M	16,5	18,5	2,2	0,21
27	F	15,2	17,5	3,1	0,35
28	F	14,7	16,3	4,2	0,47
29	F	19,8	21,5	1,0	0,30
30	M	16,2	17,2	4,9	0,46
41	F	16,7	17,7	3,8	0,37
48	F	14,9	14,1	5,2	0,44
56	M	15,3	13,9	5,9	0,61
76	F	14,1	11,7	8,2	0,84
102	M	14,2	14,1	6,7	0,75
105	M	19,3	20,5	2,1	0,19
106	M	19,1	21,5	1,6	0,18

Table 3.17 shows the parameters ($CPPS_{mean}$, HNR, Shimmer [dB] and Shimmer [%]) calculated by Matlab for THS. In addition, the trends of these parameters on the three softwares (Praat, Matlab and VOXplot) have been represented in Figure

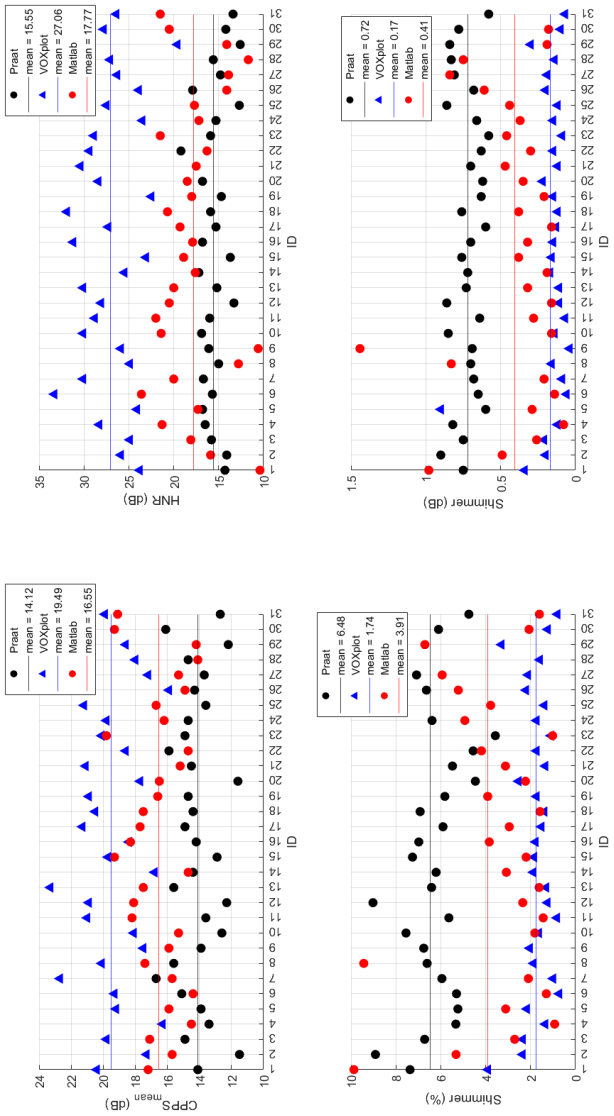

Figure 3.18: AVQI Parameters of THS data-set

3.18. As can be seen, among the software used, the one that outputs parameter values closest to the literature values is VOXplot: in fact, according to previous

studies [23] for dysphonic patients the average AVQI value is 4.27 with a standard deviation of 2.13, whereas for the control group the average AVQI value is 1.55 with a standard deviation of 0.59.

As can be seen in Figures 3.16 and 3.17, the values of the HS and MS subjects and the values of the THS subjects are very well distinguishable. For this reason, MS and THS subjects were classified according to AVQI, WS and both in order to evaluate their performance. Obviously the classification between THS and MS followed the same steps as the classification between HS and MS (paragraph 3.4). In order to have a balanced data-set, thirteen THS subjects were randomly extracted and added to the MS subjects for classification.

Table 3.18 shows the performance of the classification of subjects according to the warning score. As can be seen, the accuracy is 87.7%, which is approximately twice as high as the classification according to WS among HS and MS subjects. The confusion matrix of classification is displayed in figure 3.19, and the ROC is displayed in figure 3.20. Within the ROC curve, the AUC (Area Under Curve) value can be observed, indicating the ability of the model to differentiate between the two classes under examination. In this instance, the AUC is 91%, indicating a high classification quality of the obtained predictions.

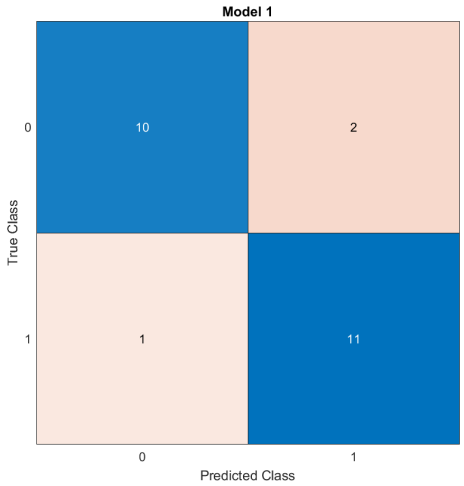

Figure 3.19: Confusion matrix of logistic regression model of Warning Score between THS and MS

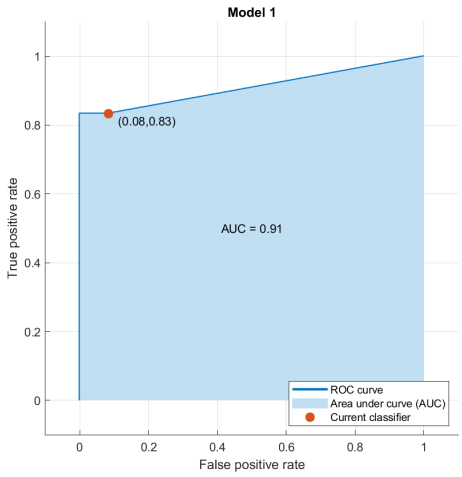

Figure 3.20: Area under ROC curve of logistic regression model of Warning Score between THS and MS

Table 3.18: Classification performance obtained for WS

THS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
Warning Score	99,0%	90,0%	83,3%	91,6%	86,9%	87,7%

Following that, THS and MS were categorized based on the AVQI that was determined using VOXplot and Praat software. Figure 3.21 displays the confusion matrix for the classification based on Praat data, whereas figure 3.23 displays the matrix based on VOXplot data.
A comparison can be made between the performance obtained by classifying THS and MS subjects according to the AVQI obtained by Praat (table 3.19) and VOXplot (table 3.20): the accuracy obtained from VOXplot data is 100%, higher than accuracy obtained from Praat that is 92.2%. A model or system that achieves 100% accuracy has correctly predicted every case or set of observations. Stated differently, it performed flawlessly when it came to predicting outcomes and categorizing data. Since there does not appear to be any margin of error for the categorization, this result might come as a surprise, however employing a larger data set would undoubtedly change it.
AUC, as shown in Figures 3.22 and 3.24, reaches a maximum value of 100% for the classification performed using VOXplot values, while it reaches a value of 99% for the classification done using Praat data.

Table 3.19: Classification performance obtained for AVQI by Praat application for THS and MS

THS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
AVQI	99,0%	92,3%	92,3%	92,3%	92,3%	92,3%

Table 3.20: Classification performance obtained for AVQI by VOXplot application of THS and MS

THS(0) vs MS(1)	AUC	Precision	Sensitivity	Specificity	F1-score	Accuracy
AVQI	100%	100%	100%	100%	100%	100%

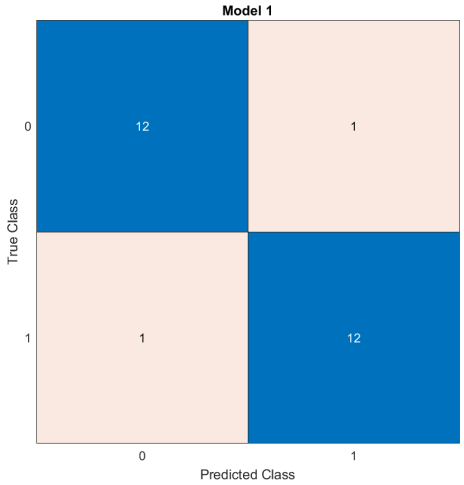

Figure 3.21: Confusion matrix of logistic regression model of Praat AVQI of THS and MS

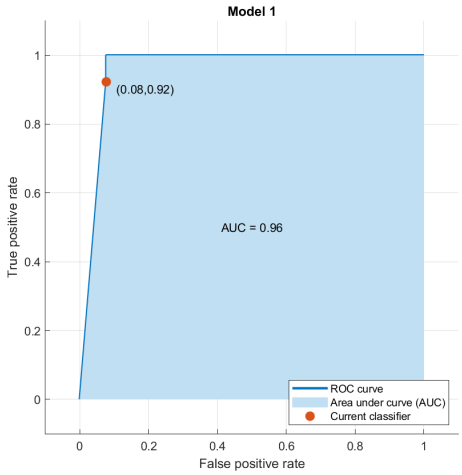

Figure 3.22: Area under ROC curve of logistic regression model of Praat AVQI THS and MS

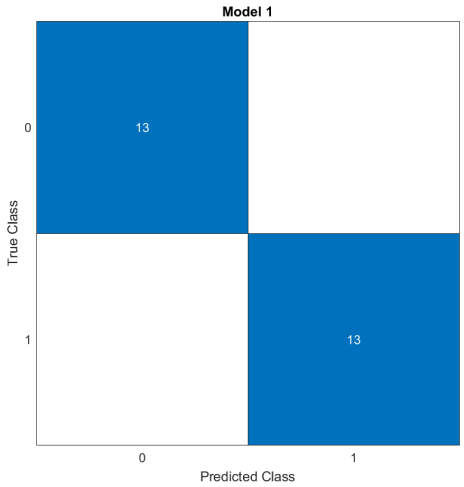

Figure 3.23: Confusion matrix of logistic regression model of VOXplot AVQI of THS and MS

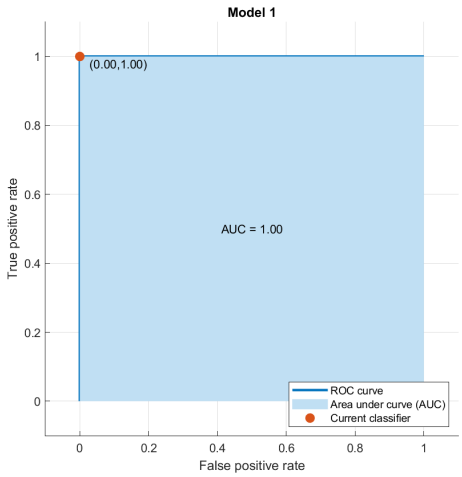

Figure 3.24: Area under ROC curve of logistic regression model of VOXplot AVQI of THS and MS

Finally, since VOXplot produced better results, we attempted to classify THS and MS using both the WS and the AVQI features. Table 3.21 displays the results. It can be seen that the accuracy gets better when compared to the WS classification, but it gets worse when compared to the VOXplot data classification. The confusion matrix is displayed in Figure 3.25 as well. The area under the

ROC curves, which reaches the lowest value when compared to the other prior classifications, is displayed in Figure 3.26 and has a value of 92%.

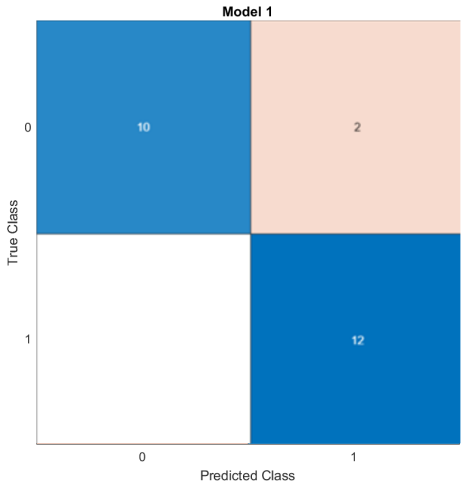

Figure 3.25: Confusion matrix of logistic regression model of VOXplot AVQI and WS of THS and MS

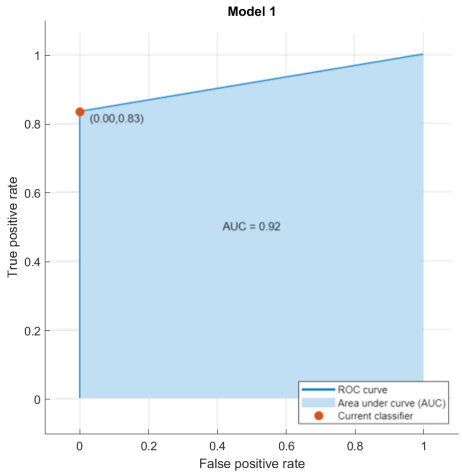

Figure 3.26: Area under ROC curve of logistic regression model of VOXplot AVQI and WS of THS and MS

Table 3.21: Classification performance obtained for AVQI by VOXplot application and WS of THS and MS

THS(0) vs MS(1)	**AUC**	**Precision**	**Sensitivity**	**Specificity**	**F1-score**	**Accuracy**
AVQI	92%	100%	83.3 %	100 %	90.9%	91.7%

Chapter 4

Conclusions

For this thesis project, different analyses on voice parameters extracted from 16 healthy subjects (HS) and 16 pathological patients with Multiple Sclerosis (MS) were carried out. For each subjects, speech material includes three vocal tasks (vocalizing the sustained vowel /a/, reading a phonetically balanced text, and giving a roughly one-minute speech). These tasks are acquired simultaneously using a contact microphone-based device named Vocal Holter (VH) and an in-air microphone system (MI). The voice parameters in the time, frequency and cepstrum domains are extracted from the harmonic frames. For the MI signal a pre-processing was done: voiced and silence selection is performed first based on intensity, but among voice frames there are some considered unharmonic including those related to the 'voiceless consonant', so a selection is made between harmonic and unharmonic frames based on HNR parameter and the frequence jump. All parameters are extracted using the same algorithm to make the values comparable. The parameters extracted are those that ,in previous studies [32], have proven to be the ones that best discriminate between HS and MS subjects and they are: for recording of sustained vowel /a/ CPPS, f0, HNR, Vam, APQ, Shimmer (db), Shimmer (%), Vfo, PPQ, RAP, local Jitter; for recording of free speech and reading task HNR and CPPS. After the extraction of the parameters, feature selection was carried out to objectively assess voice quality of HS and MS data-sets, in particular using two indices: Acoustic Voice Quality Index (AVQI) and Warning Score (WS). Jitter, shimmer, CPPS, HNR, Spectral Slope, and Tilt are the parameters that

determine AVQI. These parameters are retrieved by concatenating three seconds of recording of sustained vowel /a/ and three seconds of recording of reading task. They were extracted with different software applications: Matlab(R2022b), VOXplot and Praat. The best results were obtained with VOXplot application, in fact they were closer to the results shown in literature. In fact from literature [23], the mean AVQI value for dysphonic patients was expected to be 4.3, while for the control group the mean AVQI value was 1.6 while with VOXplot the mean value of AVQI obtained for HS subjects is 3.9, and for MS the mean value is 5.4. On the other hand, the average AVQI value obtained with Praat is 5.3 for HS, whereas for MS the average value is 6.2. Then subjects were classified according to their AVQI : subjects are divided into two classes, HS and MS, and AVQI is used to validate the LR model with 5-fold cross-validation, in the Classification Learner App in Matlab (R2022b). The best classification results were obtained with VOXplot, in fact an accuracy of 70.8% and AUC of 63% was achieved. Then the Warning Score (WS) was calculated: it depends on the parameters local jitter, local shimmer, mean and standard deviation of CPPS extracted using Matlab scripts from the vowel /a/ by both MI and VH. Subjects were also classified according to WS, obtaining an accuracy of 41.7% and an AUC of 36%; these results show a clear inconsistency with what might be expected, which is why it was assumed not to use HS as a control group, since HS are not affected by multiple sclerosis but are certainly not phonatory healthy. For this reason, an additional data-set of 57 true healthy subjects (THS), without phonatory pathologies certified, was introduced. The calculation of the parameters and the two indices, AVQI and WS, were similarly repeated. For the AVQI, only 13 THS were extracted at random in order to have a data-set balanced with the MS. Again, the best classification results, with 100% accuracy, were obtained using data from VOXplot. For WS, the classification results improved on the previous case, with an accuracy of 87.5% and an AUC of 91%. Finally, since very high accuracy values and AUC for the THS have been obtained, an attempt has been made to classify according to the two indices jointly, but no improvement has been obtained: in particular, an accuracy of 91.7% has been obtained, more than the WS but less than the one with only AVQI.

From a perspective of a bigger future project, this study may be carried out with a bigger data-set, which would undoubtedly produce better classification outcomes.To

expand these studies to a larger data-set, the speech therapists of Don Gnocchi Foundation is currently gathering more data. Furthermore, new voice signal collection technology, such as air and contact microphones that can capture signals with greater clarity and less noise and saturation, could also enhance the results. Additionally, more research might be done on the usage of software programs (such Praat and VOXplot), particularly finishing the AVQI analysis and the Spectral Slope and Tilt parameters using Matlab. To make the software consistent, another option is to investigate a way for defining reference parameters. For instance, a developer could create synthetic sample files with known parameters.

Bibliography

[1] Voice Biometrics Technologies and Applications for Healthcare: an overview- Scientific Figure on ResearchGate. accessed 12 Nov, 2022. url:https://www.researchgate.net/figure/The-peripheral-phonation-system-Purves-2012 fig2 329059822 (cit. on p. 3).

[2] Laryngeal Cartilages. url: https://teachmeanatomy.info/neck/viscera/larynx/laryngealcartilages/ (cit. on pp. 3, 4). Purves-2012 fig2 329059822 (cit. on p. 3).

[3] 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 266.11 (2009), pp. 1759–1765 (cit. on p. 5).

[4] https://www.ilmicrofono.it/content/42-l-intellegibilita-del-parlato

[5] https://www.aism.it/sintomi della sclerosi multipla disturbi del linguaggio

[6] https://www.nationalmssociety.org/What-is-MS/Types-of-MS/Relapsing-remitting-MS

[7] https://www.nationalmssociety.org/What-is-MS/Types-of-MS/Secondary-progressive-MS

[8] https://www.nationalmssociety.org/What-is-MS/Types-of-MS/Primary-progressive-MS

[9] Tanzariello. Valutazione Percettiva della Voce. url: http://www.tanzariello.it/index.php/gola/92-studio-prof-a-tanzariello/laringe/ esami/724-valutazione-percettiva-della-voce. Published: Mercoledì, 28 Marzo 2012 (cit. on p. 7).

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

[11] 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).

[12] KayPENTAX, Multi-Dimensional Voice Program (MDVP)

[13] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, IFA 17 proceedings, pp. 97-110, 1993.

[14] J. Hillenbrand, A. Houde Acoustic correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech, J. Speech Hearing Res., vol. 39, no. 2, pp. 311-321, 1996.

[15] J. Hillenbrand, R. A. Cleverand and R. L. Erickson, Acoustic correlates of breathy vocal quality, J. Speech Hearing Res., vol. 37, no. 4, pp. 769-778, 1994.

[16] Barsties B, De Bodt M. Assesment of voice quality: current state-of-the-art. *Auris Nasus Larynx*. 2015;42:183-188.

[17] Maryn Y, Roy N, De Bodt M, et al. Acoustic Measurament of overall voice quality: a meta analysis. J Acoust Soc Am. 2009;126:2619-2634.

[18] Maryn Y, Corthals P, Van Cauwenberge P, et al. Toward improved ecological validity in the acoustic measurament of overall voice quality: combining continuous speech and sustained vowels. J voice. 2010;24:540-555.

[19] Maryn Y., Corthals P et al., Toward Improved ecological Validity in the Acoustic measureament of Overall Voice Quality: Combining Continuous Speech and Sustained Vowels, Journal of voice, Vol 24, No.5, pp 540-555,2008

[20] Brown, L. M., & Davis, P. R. (Year). "Clinical applications of the Acoustic Voice Quality Index (AVQI): A review of recent studies." Journal of Speech and Hearing Research

[21] Wilson, C. R., & White, S. M. (Year). "The use of AVQI in vocal therapy: Monitoring progress and treatment outcomes." Journal of Voice

[22] Garcia, A. B., & Martinez, E. R. (Year). "Exploring the potential of AVQI in vocal research: A comprehensive analysis of vocal quality." Journal of Phonetics.

[23] Fantini M., Ricci Maccarini A., et al. Validation of the Acoustic Voice Quality Index (AVQI) Version 03.01 in Italian. 2021

[24] Paul Boersma & David Weenink (1992–2022): Praat: doing phonetics by computer [Computer program].Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

[25] Ben Barsties v. Latoszek, Jörg Mayer, Christopher R. Watts and Bernhard Lehnert. "Advances in Clinical Voice Quality Analysis with VOXplot", Advances in Clinical Voice Quality, Analysis with VOXplot. J. Clin. Med. 2023, 12, 4644. https://doi.org/10.3390/jcm12144644

[26] Astolfi A,Carullo A, Puglisi G, Fissore V, Clerici J. An experimental approach for the evaluation of vocal behaviour of university professors. Forum Acoustic 2023

[27] Tech. rep. available online. url: https://towardsdatascience.com/aquick-guide-to-auc-roc-in-machine-learning-models-f0aedb78fbad (cit. on p. 35).

[28] Daniel Jurafsky & James H. Martin. "Logistic Regression", Speech and Language Processing, 2023

[29] Dalibor Mitrović,Matthias Zeppelzauer, Christian Breiteneder, Advances in Computers, Volume 78, 2010, Pages 71-15

[30] Allan D. Pierce. Acoustics: An Introduction to Its Physical Principles and Applications, 2019.

[31] https://www.vvl.be/documenten-en-paginas/praat-script-avqi-v0203

[32] A. Fantoni. Assessment of Vocal Fatigue of Multiple Sclerosis Patients - Validation of a Contact Microphone-based Device for Long-Term Monitoring. Politecnico di Torino, 2023

[33] Hiroya Yamaguchi, Rahul Shrivastav, Moya L Andrews, Seiji Niimi. A comparison of voice quality ratings made by Japanese and American listeners using the GRBAS scale. 2003 May-Jun;55(3):147-57

[34] Sara Palmieri. «Assessment of the vocal status of multiple sclerosis patients comparison with healthy subjects and evaluation of vocal rehabilitation». MsC thesis. Politecnico di Torino, 2023

Appendix A

Praat Script

TITLE OF THE SCRIPT: ACOUSTIC VOICE QUALITY INDEX (AVQI) v.03.01

Form for introduction and/or parameterization form Acoustic Voice Quality Index v.03.01 It is advocated to estimate someone's dysphonia severity in both continuous speech (i.e., 'cs') and sustained vowel (i.e., 'sv') (Maryn et al.,comment 2010). This script therefore runs on these two types of recordings, and it is important to name these recordings 'cs' and 'sv', respectively. This script automatically (a) searches, extracts and then concatenates the voiced segments of the continuous speech recording to a new sound; (b) concatenates the sustained vowel recording to the new sound, (c) determines the Smoothed Cepstral Peak Prominence, the Shimmer Local, the Shimmer Local dB, the LTAS-slope, the LTAS-tilt and the Harmonics-to-Noise Ratio of the concatenated sound signal, (d) calculates the AVQI-score based on the equation of Barsties & Maryn (2015), and draws the oscillogram, the narrow-band spectrogram with LTAS and the power-cepstrogram with power-cepstrum of the concatenated sound signal to allow further interpretation. To be reliable for the AVQI analysis, it is imperative that the sound recordings are made in an optimal data acquisition conditions. There are two versions in this script: (1) a simple version (only AVQI with data of acoustic measures), and (2) an illustrated version (AVQI with data of acoustic measures and above-mentioned graphs).

choice version: 1
button simple
button illustrated

comment »> Additional information (optional):
sentence namepatient
sentence leftdates(birth-assessment)
sentence rightdates(birth-assessment)
comment
comment Script credits: Youri Maryn (PhD), Paul Corthals (PhD), and Ben Barsties
endform

Erase all
Select inner viewport... 0.5 7.5 0.5 4.5
Axes... 0 1 0 1
Black
Text special... 0.5 centre 0.6 half Helvetica 12 0 Please wait an instant. Depending on the duration and/or the sample rate of the recorded
Text special... 0.5 centre 0.4 half Helvetica 12 0 sound files, this script takes more or less time to process the sound and search for the AVQI.

PART 0: HIGH-PASS FILTERING OF THE SOUND FILES.

---

select Sound cs
Filter (stop Hann band)... 0 34 0.1
Rename... cs2
select Sound sv
Filter (stop Hann band)... 0 34 0.1
Rename... sv2

PART 1: DETECTION, EXTRACTION AND CONCATENATION OF THE VOICED SEGMENTS IN THE RECORDING OF CONTINUOUS SPEECH.

---

select Sound cs2
Copy... original

```
samplingRate = Get sampling frequency
intermediateSamples = Get sampling period
Create Sound... onlyVoice 0 0.001 'samplingRate' 0
select Sound original
To TextGrid (silences)... 50 0.003 -25 0.1 0.1 silence sounding
select Sound original
plus TextGrid original
Extract intervals where... 1 no "does not contain" silence
Concatenate
select Sound chain
Rename... onlyLoud
globalPower = Get power in air
select TextGrid original
Remove


   select Sound onlyLoud
signalEnd = Get end time
windowBorderLeft = Get start time
windowWidth = 0.03
windowBorderRight = windowBorderLeft + windowWidth
globalPower = Get power in air
voicelessThreshold = globalPower*(30/100)


   select Sound onlyLoud
extremeRight = signalEnd - windowWidth
while windowBorderRight < extremeRight
Extract part... 'windowBorderLeft' 'windowBorderRight' Rectangular 1.0 no select
Sound onlyLoudpart
partialPower = Get power in air
if partialPower > voicelessThreshold
call checkZeros 0
if (zeroCrossingRate <> undefined) and (zeroCrossingRate < 3000)
select Sound onlyVoice
plus Sound onlyLoudpart
```

```
Concatenate
Rename... onlyVoiceNew
select Sound onlyVoice
Remove
select Sound onlyVoiceNew
Rename... onlyVoice
endif
endif
select Sound onlyLoudpart
Remove
windowBorderLeft = windowBorderLeft + 0.03
windowBorderRight = windowBorderLeft + 0.03
select Sound onlyLoud
endwhile
select Sound onlyVoice

    procedure checkZeros zeroCrossingRate

    start = 0.0025
startZero = Get nearest zero crossing... 'start'
findStart = startZero
findStartZeroPlusOne = startZero + intermediateSamples
startZeroPlusOne = Get nearest zero crossing... 'findStartZeroPlusOne'
zeroCrossings = 0
strips = 0

    while (findStart < 0.0275) and (findStart <> undefined)
while startZeroPlusOne = findStart
findStartZeroPlusOne = findStartZeroPlusOne + intermediateSamples
startZeroPlusOne = Get nearest zero crossing... 'findStartZeroPlusOne'
endwhile
afstand = startZeroPlusOne - startZero
strips = strips +1
zeroCrossings = zeroCrossings +1
```

```
findStart = startZeroPlusOne
endwhile
zeroCrossingRate = zeroCrossings/afstand
endproc
```

PART 2: DETERMINATION OF THE SIX ACOUSTIC MEASURES AND CALCULATION OF THE ACOUSTIC VOICE QUALITY INDEX.

```
   select Sound sv2
durationVowel = Get total duration
durationStart=durationVowel-3
if durationVowel>3
Extract part... durationStart durationVowel rectangular 1 no
Rename... sv3
elsif durationVowel<=3
Copy... sv3
endif

   select Sound onlyVoice
durationOnlyVoice = Get total duration
plus Sound sv3
Concatenate
Rename... avqi
durationAll = Get total duration
minimumSPL = Get minimum... 0 0 None
maximumSPL = Get maximum... 0 0 None

   Narrow-band spectrogram and LTAS

   To Spectrogram... 0.03 4000 0.002 20 Gaussian
select Sound avqi
To Ltas... 1
minimumSpectrum = Get minimum... 0 4000 None
```

```
maximumSpectrum = Get maximum... 0 4000 None
```

Power-cepstrogram, Cepstral peak prominence and Smoothed cepstral peak prominence

```
select Sound avqi
To PowerCepstrogram... 60 0.002 5000 50
cpps = Get CPPS... no 0.01 0.001 60 330 0.05 Parabolic 0.001 0 Straight Robust
To PowerCepstrum (slice)... 0.1
maximumCepstrum = Get peak... 60 330 None
```

Slope of the long-term average spectrum

```
select Sound avqi
To Ltas... 1
slope = Get slope... 0 1000 1000 10000 energy
```

Tilt of trendline through the long-term average spectrum

```
select Ltas avqi
Compute trend line... 1 10000
tilt = Get slope... 0 1000 1000 10000 energy
```

Amplitude perturbation measures

```
select Sound avqi
To PointProcess (periodic, cc)... 50 400
Rename... avqi1
select Sound avqi
plus PointProcess avqi1
percentShimmer = Get shimmer (local)... 0 0 0.0001 0.02 1.3 1.6
shim = percentShimmer*100
shdb = Get shimmer (localdB)... 0 0 0.0001 0.02 1.3 1.6
```

Harmonic-to-noise ratio

```
select Sound avqi
To Pitch (cc)... 0 75 15 no 0.03 0.45 0.01 0.35 0.14 600
select Sound avqi
plus Pitch avqi
To PointProcess (cc)
Rename... avqi2
select Sound avqi
plus Pitch avqi
plus PointProcess avqi2
voiceReport$ = Voice report... 0 0 75 600 1.3 1.6 0.03 0.45
hnr = extractNumber (voiceReport$, "Mean harmonics-to-noise ratio: ")
```

Calculation of the AVQI

```
avqi = (4.152-(0.177*cpps)-(0.006*hnr)-(0.037*shim)+(0.941*shdb)+
+ 0.01*slope)+(0.093*tilt))*2.8902
```

PART 3: DRAWINGS ALL THE INFORMATION AND THE GRAPHS.

Title and patient information

```
Erase all
Solid line
Line width... 1
Black
Helvetica
Select inner viewport... 0 8 0 0.5
Font size... 1
Select inner viewport... 0.5 7.5 0.1 0.15
Axes... 0 1 0 1
Text... 0 Left 0.5 Half Script: Youri Maryn (PhD) and Paul Corthals (PhD)
Font size... 12
```

```
Select inner viewport... 0.5 7.5 0 0.5
Axes... 0 1 0 1
Text... 0 Left 0.5 Half ACOUSTIC VOICE QUALITY INDEX (AVQI) v.03.01
Font size... 8
Select inner viewport... 0.5 7.5 0 0.5
Axes... 0 1 0 3
Text... 1 Right 2.3 Half %%'name_patient$'%
Text... 1 Right 1.5 Half %% 'left_dates$'%
Text... 1 Right 0.7 Half %%'right_dates$'%

   Simple version

   if version = 1

   Data

   Font size... 10
Select inner viewport... 0.5 7.5 0.5 2
Axes... 0 7 6 0
Text... 0.05 Left 0.5 Half Smoothed cepstral peak prominence (CPPS): 'cpps:2'
Text... 0.05 Left 1.5 Half Harmonics-to-noise ratio: 'hnr:2' dB
Text... 0.05 Left 2.5 Half Shimmer local: 'shim:2' %
Text... 0.05 Left 3.5 Half Shimmer local dB: 'shdb:2' dB
Text... 0.05 Left 4.5 Half Slope of LTAS: 'slope:2' dB
Text... 0.05 Left 5.5 Half Tilt of trendline through LTAS: 'tilt:2' dB
Select inner viewport... 0.5 3.8 0.5 2
Draw inner box
Font size... 7
Arrow size... 1
Select inner viewport... 4 7.5 1.25 2
Axes... 0 10 1 0
Paint rectangle... green 0 2.43 0 1
Paint rectangle... red 2.43 10 0 1
Draw arrow... avqi 1 avqi 0
```

```
Draw inner box
Marks top every... 1 1 yes yes no
Font size... 16
Select inner viewport... 4 7.5 0.5 1.15
Axes... 0 1 0 1
Text... 0.5 Centre 0.5 Half AVQI: 'avqi:2'

    Copy Praat picture

    Select inner viewport... 0.5 7.5 0 2
Copy to clipboard

    Illustrated version

    elsif version = 2

    Oscillogram

    Font size... 7
Select inner viewport... 0.5 5 0.5 2.0
select Sound avqi
Draw... 0 0 0 0 no Curve
Draw inner box
One mark left... minimumSPL no yes no 'minimumSPL:2'
One mark left... maximumSPL no yes no 'maximumSPL:2'
Text left... no Sound pressure level (Pa)
One mark bottom... 0 no yes no 0.00
One mark bottom... durationOnlyVoice no no yes
One mark bottom... durationAll no yes no 'durationAll:2'
Text bottom... no Time (s)

    Narrow-band spectrogram

    Select inner viewport... 0.5 5 2.3 3.8
```

```
select Spectrogram avqi
Paint... 0 0 0 4000 100 yes 50 6 0 no
Draw inner box
One mark left... 0 no yes no 0
One mark left... 4000 no yes no 4000
Text left... no Frequency (Hz)
One mark bottom... 0 no yes no 0.00
One mark bottom... durationOnlyVoice no no yes
One mark bottom... durationAll no yes no 'durationAll:2'
Text bottom... no Time (s)
```

LTAS

```
Select inner viewport... 5.4 7.5 2.3 3.8
select Ltas avqi
Draw... 0 4000 minimumSpectrum maximumSpectrum no Curve
Draw inner box
One mark left... minimumSpectrum no yes no 'minimumSpectrum:2'
One mark left... maximumSpectrum no yes no 'maximumSpectrum:2'
Text left... no Sound pressure level (dB/Hz)
One mark bottom... 0 no yes no 0
One mark bottom... 4000 no yes no 4000
Text bottom... no Frequency (Hz)
```

Power-cepstrogram

```
Select inner viewport... 0.5 5 4.1 5.6
select PowerCepstrogram avqi
Paint... 0 0 0.00303 0.01667 0 0 no
Draw inner box
One mark left... 0.00303 no yes no 0.003
One mark left... 0.01667 no yes no 0.017
Text left... no Quefrency (s)
One mark bottom... 0 no yes no 0.00
```

```
One mark bottom... durationOnlyVoice no no yes
One mark bottom... durationAll no yes no 'durationAll:2'
Text bottom... no Time (s)

    Power-cepstrum

    Select inner viewport... 5.4 7.5 4.1 5.6
select PowerCepstrum avqi_0_100
Draw... 0.00303 0.01667 0 0 no
Draw tilt line... 0.00303 0.01667 0 0 0.00303 0.01667 Straight Robust
Draw inner box
One mark left... maximumCepstrum no yes no 'maximumCepstrum:2'
Text left... no Amplitude (dB)
One mark bottom... 0.00303 no yes no 0.003
One mark bottom... 0.01667 no yes no 0.017
Text bottom... no Quefrency (s)

    Data

    Font size... 10
Select inner viewport... 0.5 7.5 5.9 7.4
Axes... 0 7 6 0
Text... 0.05 Left 0.5 Half Smoothed cepstral peak prominence (CPPS): 'cpps:2'
Text... 0.05 Left 1.5 Half Harmonics-to-noise ratio: 'hnr:2' dB
Text... 0.05 Left 2.5 Half Shimmer local: 'shim:2' %
Text... 0.05 Left 3.5 Half Shimmer local dB: 'shdb:2' dB
Text... 0.05 Left 4.5 Half Slope of LTAS: 'slope:2' dB
Text... 0.05 Left 5.5 Half Tilt of trendline through LTAS: 'tilt:2' dB
Select inner viewport... 0.5 3.8 5.9 7.4
Draw inner box
Font size... 7
Arrow size... 1
Select inner viewport... 4 7.5 6.75 7.4
Axes... 0 10 1 0
```

```
Paint rectangle... green 0 2.43 0 1
Paint rectangle... red 2.43 10 0 1
Draw arrow... avqi 1 avqi 0
Draw inner box
Marks top every... 1 1 yes yes no
Font size... 16
Select inner viewport... 4 7.5 5.9 6.65
Axes... 0 1 0 1
Text... 0.5 Centre 0.5 Half AVQI: 'avqi:2'

    Copy Praat picture

    Select inner viewport... 0.5 7.5 0 7.4
Copy to clipboard

    endif

    Remove intermediate objects

    select all
minus Sound sv
minus Sound cs
minus Sound avqi
Remove
```