POLYTECHNIC OF TURIN

Master's Degree in Biomedical Engineering

Master's Degree Thesis

Assessment of the vocal status of multiple sclerosis patients

Comparison with healthy subjects and evaluation of vocal rehabilitation

Supervisor

Prof. Alessio CARULLO

Prof. Alberto VALLAN

Candidate

Sara PALMIERI

MARCH 2023

Summary

Multiple Sclerosis (MS) is an autoimmune disease that involves the central nervous system due to a loss of myelin, the substance that enables neurons to transmit electrical signals between the brain and the periphery of the body. Consequently, all body activities of MS patients are impaired, including speech articulation. This study was carried out in collaboration with the team of speech therapists at the Don Gnocchi Foundation in Milan, analysing the speech performance of about 70 patients that suffer from MS. One group pf patients was treated with standard therapy, a second group with innovative LSVT-LOUD therapy and a third group was not treated. The recordings provided include 3 repetitions of the vowel /a/ and a free speech (about 1 minute), which were acquired for each patient before and after therapy. After excluding not valid recordings (saturated or too noisy), two subsets were created for the analysis of the /a/ and the free speech. Using scripts developed in the Matlab R2020a environment, 9 descriptive statistics of Harmonic-to-Noise Ratio (HNR), intensity (RMS), Cepstral Peak Prominence Smoothed (CPPS) and fundamental frequency (fo) were extracted. In the case of the vowel /a/ other 9 amplitude and period stability parameters were obtained. The script first performs a pre-processing that selects non-silent harmonic signal frames with frequency jumps between adjacent frames not greater than half octave. For the vowel /a/ it was necessary to carry out a further cleaning: pre-processing sometimes resulted in the elimination of internal frames, thus creating artefacts in the calculation of stability parameters. To solve this problem, only adjacent frames were kept. The parameters of the recordings of 60 healthy subjects, which are available at Politecnico di Torino, were extracted with the same script. By comparing the vocal parameters of SM patients before therapy (T_0) with those of healthy subjects, the most distinguishable values have been identified. Parameters

identified for the vowel /a/: stability parameters of amplitude and period, 5° percentile (prc) and standard deviation (std) of CPPS, 95° prc, range and std of fo, mean, median, mode and 5° prc of HNR. Parameters identified for the free speech: mean, median, mode, std, range, 5° prc, 95° prc, skewness of CPPS, std and range of HNR. In addition, comparing the values of the patients pre and post therapy, the parameters mainly affected by the therapy have been observed. Then, SM patients were analysed observing the difference between the parameters extracted at T_1 (post-therapy) and $T0$ (pre-therapy). The patients were divided into three classes according to the therapy (LSVT-LOUD, ACTIVE, no therapy) and the most representative features were sought for distinguishing the therapies. For this aim, a combinatorial algorithm based on the logistic regression model was used. Taking two classes at time, the model was tested with single features and with all possible combinations of 2, 3 and 4 features and selecting those that exhibited the best classification performance. The same operation was performed by weighting the features with the reciprocal and the complement to one of the std of the three repeated vowels, but no relevant improvements were observed. Subsequently, the best feature combinations were validated (5-fold cross-validation) through the Matlab APP Classification Learner. The best performance were obtained for the vowel /a/ using the stability parameters vfo, apq Vam and Jitter% and the statistics 95° prc, 5° prc, range, mode, median of HNR and std, 95° prc of fo, range, std, 95° prc, kurtosis of the RMS; for free speech using the statistics mean, mode, std, range, 5° prc, 95° prc of CPPS, mean, median, mode, range of HNR, 95° prc, std of fo, mode, 5° prc for the RMS. The best validated accuracy was of about 82% for free speech. The consistency between the perceptual evaluation of the experts, which were provided using the G value of the GIRBAS scale, and the obtained outcomes was assessed. The differences in G values between T_1 and T_0 were compared to the feature differences, reporting the results as a confusion matrix and taking the experts' assessment as a reference. From this last analysis, many errors were observed that are mainly related to the poor resolution of the GIRBAS scale (0 to 4).

Acknowledgements

This study was the result of a collaboration between the Politecnico di Torino and the team of speech therapists from the Don Gnocchi Foundation in Milan, who made available the voice recordings acquired before and after voice therapy of around 70 multiple sclerosis patients.

Table of Contents

List of Tables VIII

List of Figures X

1 The speech 1

1.1 Anatomy and Physiology of the Voice Production . . . 2

1.1.1 Air Pressure System . . . 2

1.1.2 Phonatory System . . . 2

1.1.3 Articulatory System . . . 3

1.2 Voice Signal . . . 4

1.3 Vocal symptoms and acoustic changes in patients with multiple sclerosis . . . 5

1.4 Voice Rehabilitation Technique . . . 6

1.4.1 The Lee Silverman Voice Treatment LOUD . . . 6

1.4.2 LSVT-LOUD versus Standard therapy . . . 6

1.5 Perceptual Rating Scales: GIRBAS and VHI . . . 7

2 Materials and Methods 9

2.1 Data-set . . . 10

2.2 Signal Processing . . . 10

2.2.1 Manual Cleaning . . . 10

2.2.2 Pre-processing . . . 12

2.2.3 Features Extraction . . . 13

2.2.4 Acoustic Parameters . . . 14

2.2.5 Recording Parameters 17
2.2.6 Stability Parameters 18
2.3 Data cleaning (vowel /a/) 22
2.4 Pathological vs. Healthy subjects 23
2.5 Feature Selection 26
2.5.1 Data Observation 26
2.5.2 Logistic Regression 30
2.5.3 Feature Selection using Logistic Regression 36
2.5.4 Feature Validation 39
2.6 Relation between VHI and GIRBAS scales 39
2.7 Comparison between Extracted Features and GIRBAS scale 41

3 Results and Discussion 48
3.1 Pathological vs. Healty Results 48
3.1.1 Vowel /a/ Results 49
3.1.2 Free speech Results 51
3.2 Logistic Regression Result 51
3.2.1 Vowel /a/ Results 52
3.2.2 Free speech Results 60
3.3 VHI vs. GIRBAS scale Results 64
3.4 Extracted Features vs. GIRBAS scale Results 64
3.4.1 Vowel /a/ Results 65
3.4.2 Free speech Results 68

4 Conclusions 70

Bibliography 72

List of Tables

2.1 Data-set for the analysis of the vowel /a/, patients are divided in 3 therapy classes with the corresponding numerosity. 12

2.2 Data-set for the analysis of the free speech, patients are divided in 3 therapy classes with the corresponding numerosity. 12

2.3 The 3 class combinations for binary classification with their respective labels, the two columns on the right show the class numerosity for the two analyses performed. 36

3.1 Average accuracy of the relationship between the G parameter (GIRBAS scales) and extracted parameters. 65

3.2 Average accuracy of the relationship between the A GIRBAS parameter (scales) and extracted parameters. 65

3.3 Average precision of the relationship between the G parameter (GIRBAS scale) and extracted parameters. 66

3.4 Average precision of the relationship between the A parameter (GIRBAS scale) and extracted parameters. 66

3.5 Average sensitivity of the relationship between the G parameter (GIRBAS scale) and extracted parameters. 66

3.6 Average sensitivity of the relationship between the A parameter (GIRBAS scale) and extracted parameters. 66

3.7 Average specificity of the relationship between the G parameter (GIRBAS scale) and extracted parameters. 67

3.8 Average specificity of the relationship between the A parameter (GIRBAS scale) and extracted parameters. 67

3.9 Evaluation metrics averaged over all features, in relation to parameter G, for all patients under examination. 67
3.10 Evaluation metrics averaged over all features, in relation to parameter A, for all patients under examination. 67
3.11 Average accuracy of the relationship between the G parameter (GIR-BAS scales) and extracted parameters. 68
3.12 Average accuracy of the relationship between the A parameter (GIR-BAS scales) and extracted parameters. 68
3.13 Average precision of the relationship between the G parameter (GIR-BAS scale) and extracted parameters. 68
3.14 Average precision of the relationship between the A parameter (GIR-BAS scale) and extracted parameters. 68
3.15 Average sensitivity of the relationship between the G parameter (GIRBAS scale) and extracted parameters. 69
3.16 Average sensitivity of the relationship between the A parameter (GIRBAS scale) and extracted parameters. 69
3.17 Average specificity of the relationship between the G parameter (GIRBAS scale) and extracted parameters. 69
3.18 Average specificity of the relationship between the A parameter (GIRBAS scale) and extracted parameters. 69
3.19 Evaluation metrics averaged over all features, in relation to parameter G, for all patients under examination. 69
3.20 Evaluation metrics averaged over all features, in relation to parameter A, for all patients under examination. 69

List of Figures

1.1 Generic representation of the voice production apparatus [2]. 3

1.2 Open and close vocal folds [3]. 4

1.3 Vocal signal during the phonation of the word "sì" (with Italian pronunciation) extracted from Audacity. 5

2.1 Problems encountered during manual pre-processing: a) instrumental artifacts b) low frequency artifact c) high frequency artifact d) saturation 12

2.2 In blue the cepstrum extracted from a patient speech recording, in red the related regression line, the quefrency in which the peak falls corresponds to the fundamental period of the signal. 17

2.3 Two examples relating to the recording of the vowel /a/ sustained, on the x's are the total frames on the y's are associated the value: 1) harmonic frame, 2) non-harmonic and 0) silence. 23

2.4 Absolute value of jitter % of pathological patients (yellow to red as the severity of vocal ability increases) and healthy subjects in blue . 24

2.5 Average value and dispersion ($\pm 2\sigma$) of jitter with reference to the 4 categories under analysis. 25

2.6 Mean value over the 3 repetitions of the vowel /a/ of the mean HNR over the two time observations T_0 and T_1 for the patients in the analysis divided by class. 27

2.7 Vowel /a/: Mean HNR delta values ($T_1 - T_0$) and relative standard deviations of the subjects divided into the three classes. 28

2.8 Free speech: Mean CPPS delta values ($T_1 - T_0$) of the subjects divided into the three classes. 29
2.9 Vowel /a/: Average feature values with relative dispersion considering the range $\pm\sigma$, on the abscissas are the classes with the total number of patients belonging to them. 30
2.10 Free speech: Average feature values with relative dispersion considering the range $\pm\sigma$, on the abscissas are the classes with the total number of patients belonging to them. 31
2.11 Single-variate Logistic Regression model. 32
2.12 Confusion Matrix in binary classification. 33
2.13 Examples of ROC curves of different classifiers [12]. 35
2.14 Scatter plot comparing self-reported VHI and perceptual G values (Girbas scale) of patients. 40
2.15 Scatter plot comparing self-reported VHI and perceptual A values (girbAs scale) of patients. 41
2.16 Relationship between features and component G of the GIRBAS assessment, in both cases positive values correspond to an improvement following therapy. 42
2.17 Relationship between features and component A of the GIRBAS assessment, in both cases positive values correspond to an improvement following therapy. 43
2.18 CM general (all patients) and partial (divided by therapy classes) where the true class is obtained from the GIRBAS (only G component) evaluation and the predicted is obtained from parameter extraction. 44
2.19 CM general (all patients) and partial (divided by therapy classes) where the true class is obtained from the GIRBAS (only A component) evaluation and the predicted is obtained from parameter extraction. 45
2.20 Ideal relation between delta figures and delta G. 47
2.21 Ideal relation between delta figures and delta A. 47

3.1 The table shows the 17 parameters that are statistically independent (95.45%) when comparing MS patients (T_0) vs. Healthy 50
3.2 The table shows the 4 parameters that are statistically independent (95.45%) when comparing MS patients (T_0) vs. Healthy 52
3.3 Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 53
3.4 Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 54
3.5 Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 54
3.6 validated classification results without features selection 55
3.7 Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 56
3.8 Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 56
3.9 Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 57

3.10 validated classification results without features selection 57
3.11 Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 58
3.12 Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 59
3.13 Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 60
3.14 validated classification results without features selection 60
3.15 Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 61
3.16 validated classification results without features selection 62
3.17 Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 62
3.18 Validated classification results without features selection 63
3.19 Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection 63
3.20 validated classification results without features selection 64

Chapter 1

The speech

"We talk to communicate, to feel less alone. Our bodies were designed for this. Our speech is an important aspect of who we are. There are people who spit words, who strike them like matches against our wrists. There are people who put us back together, who slowly stack our limbs upon each other with a few simple words. There are those who say nothing. There are those who say too much. In special cases there are those who let you speak, then act as if you never said a thing."
-Liv Baker

This chapter briefly explains the physiological systems involved in voice and speech production. An introduction to the speech signal and its main characteristics is also made. In particular, the attention is put on the phonation difficulties of multiple sclerosis patients and the related vocal rehabilitation techniques; two rehabilitation techniques are discussed, one standard currently used by speech therapists with MS patients, and a second technique called LSVT-LOUD used so far for the treatment of Parkinson's patients. Eventually, perceptual and self-assessment scales to classify vocal abilities are presented.

1.1 Anatomy and Physiology of the Voice Production

The voice production is a complex process that includes several anatomic components, it could be divided in three system with specific roles as illustrated in [1]. The first regulates the airflow which is the power supply for the voice (air pressure system). The air interacts with specific anatomic structures and makes them vibrate (phonatory system). That vibration produces an acoustic wave that is modulated and propagates outwards through the oral cavity (articulatory system). Fig 1.1 shows a generic representation of the voice production apparatus and the anatomic structures involved.

1.1.1 Air Pressure System

The air pressure system is composed by the diaphragm, abdominal and chest muscles, lungs, ribs and trachea. The process starts with inspiration, the air goes trough the oral and nasal cavity, pass the trachea and arrives in the lungs. The diaphragm flat down and let the ribcage expand so the air can go inside the lungs. Once the lungs reach capacity the lungs elastic tissue recoils and the air is exhaled. The exhaled air goes up to the trachea and pass trough the larynx where interacts with the vocal fold.

1.1.2 Phonatory System

The phonatory system converts the air flow energy in acoustic energy trough the interaction between the air flow and the vocal folds (Fig:1.2) located in the larynx. The larynx is a structure with a cartilaginous support that ends superiorly with the hyoid bone and inferiorly with the trachea. The glottis is the space between the vocal folds, the glottis closure generates a resistance to the air that comes from the lungs. When the pressure of the air overcomes the glottis's closing force the vocal folds are separated. Once the glottis is completely open the air pressure decreases and the vocal folds re-approximates. This phenomena is called the "vibratory cycles" and the number of repetitions of this cycle per second determines the acoustic wave frequency. The acoustic waves at the specific frequency are propagated to the

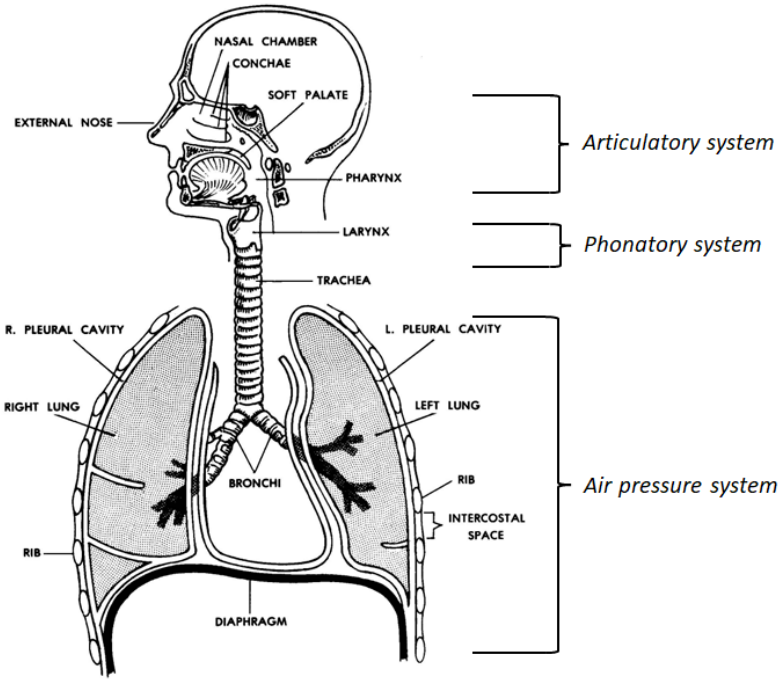

Figure 1.1: Generic representation of the voice production apparatus [2].

articulatory system.

1.1.3 Articulatory System

The waves, produced by the phonatory system, arrives to the oral and nasal cavities that filter the acoustic waves until it emerges from the mouth and nostrils. The cavities are resonant environments in which components far from the resonance are mitigated while those close to it are amplified. This phenomenon creates secondary frequencies called formants that modulate the course of the original wave and allow the shape of the voice to be adjusted. The oral cavity can be modulated by the movement of the tongue, jaw and lips allowing the articulation of speech. This modulation, in the case of consonants, creates constrictions in which air is forced to pass or is stopped; this generates turbulent air movements that add to the periodic wave obtained from the vibration of the vocal cords.

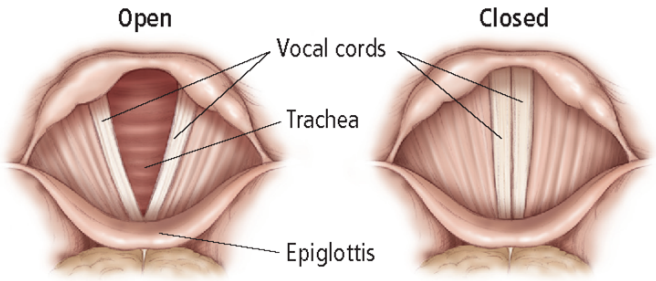

Figure 1.2: Open and close vocal folds [3].

1.2 Voice Signal

The speech signal is a complex signal in the time domain consisting of a harmonic and a non-harmonic part. It can be studied in the time domain, where the waveform can be observed, or in the frequency domain, where harmonic components (peaks centred in the characteristic frequency) and non-harmonic components (constant white noise component) can be distinguished. In general, two types of sounds are identified in vocal production:

- Vocalized sounds: sounds generated by the passage of air through the glottis and thus by the vibration of the vocal cords. These sounds are characterized by the fundamental frequency, determined by the opening and closing of the vocal cords, and other frequencies, called formants, generate by the resonant cavities.

- Voiceless sounds: sounds that do not involve the vocal folds, air is forced through a bottleneck in the oral tract to create turbulence. An example would be consonants /s/, /f/ and /t/ in the Italian language.

Observing Fig:1.3, vocal signal during the phonation of the word "sì" (with Italian pronunciation), it is possible to recognise that in the first tract (voiceless consonant /s/) the signal is similar to a statistic noise instead in the second tract (vocalized vowel /i/) the signal is more similar to a sinusoidal wave. A signal can be analyzed in the time domain, as in the last example, or in the frequency domain, where it is possible to search for fundamental frequencies and the formants. The fundamental frequency of the vocalized sounds is different for men and women, for the first is

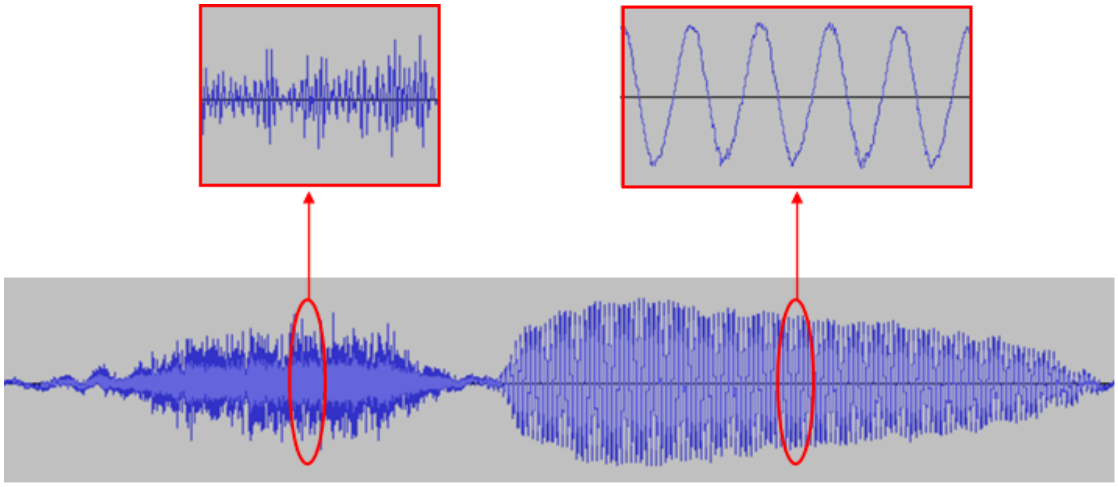

Figure 1.3: Vocal signal during the phonation of the word "sì" (with Italian pronunciation) extracted from Audacity.

from 75 Hz to 300 Hz and for the seconds is from 100 Hz to 400 Hz; this is due to the different conformation of the vocal folds which are wider and longer in the men case.

1.3 Vocal symptoms and acoustic changes in patients with multiple sclerosis

Speech is a complex activity that requires the coordination of different systems in the body including the neurologic system and is a reflection of the well-being of the entire body [4]. The multiple sclerosis (MS) is a chronic degenerative disease that affects the myelin sheath creating multiples lesion on the brain white matter, brainstem and spinal cord leading to considerable physical disability. The laryngeal pathophysiology reflects neuromuscular disorders with dysphonia as one of the first presenting symptoms. A physician with a great experience may deduce relevant clinical information by the phonatory characteristic of the patient. Unfortunately patient with MS have sometimes intermittent or subtle presence of vocal changes so often they are not perceived. Acoustic analysis can assist the physician's perceptual evaluation and support expert assessments.

1.4 Voice Rehabilitation Technique

Referring to [5] at least 62% of people affected by multiple sclerosis have vocal disorders. Vocal symptoms could be very different due to the different problems including respiratory muscles weakness, incomplete glottal closure and posterior glottal chink. The predominant vocal issues are phonatory instability, vocal asthenia, harshness, breathiness and hypophonia. Hypophonia is the most common symptoms (present in 16% of patients with MS) occurring in the early stages of the disease. Vocal disorders affects significantly the quality of life but just the 2% of the patients with MS (PwMS) receive speech therapy and few studies have been conducted on the effectiveness of vocal treatments.

1.4.1 The Lee Silverman Voice Treatment LOUD

The Lee Silverman Voice Treatment LOUD (LSVT-LOUD), as explained in the [6], is a popular voice therapy normally used for treat people with Parkinson's disease (PD), in fact the leading cause of death in patients with PD is aspiration pneumonia associated with the presence of dysphagia and dystussia. LSVT-LOUD primarily aims to improve the loudness but at the same time it effects the sensomotor speech system with improvements in speech intelligibility, facial expression, breath support and voice quality. Each session includes both "daily tasks" that are always the same and "hierarchical exercises". Daily tasks comprise 30 minutes of sustained phonation of the vowel /a/, high volume /a/, pitch glides and 10 functional sentences; instead hieranchical exercises includes 30 minutes reading exercises with increasing difficulties in duration and complexity of the tasks. Unlike other vocal treatment, this technique requires intensive high effort speech exercise during treatment combined with a simple and continuous exercise during every day life.

1.4.2 LSVT-LOUD versus Standard therapy

The standard therapy includes all speech techniques among which are exercises targeting respiration, phonation and behavioral strategies. The exercises are adapted and personalized on patient needs. In particular the rehabilitation protocol comprehended tree type of exercises:

1. Increase diaphragmatic and respiratory function and improve respiratory flow awareness;

2. Improve phonatory stability and glottis closure increasing resistance to air passage;

3. Increase expiration time and pneumo-phono-articulatory coordination with exercises of growing difficulties.

The substantial difference beetwen this two type of vocal therapy lies in the fact that LSVT therapy follows a standard protocol while conventional therapy is defined by the therapist adapting it to the patient and thus aims for targeted and personalized improvement.

1.5 Perceptual Rating Scales: GIRBAS and VHI

Patients' pathological voices need to be classified to assess the severity of the vocal problems, to observe any changes following injuries, surgery and rehabilitation therapy but also to simplifies communication between therapists. GIRBAS is a perceptive evaluation scale [7] that includes 6 parameters assigned by the therapists:

- Grade: generic grade of dysphonia;

- Instability: instability of voice functionality over time, very important for long term evaluation;

- Roughness: low frequency aperiodicity related to atypically vocal folds vibration, that generates fluctuation in wave fundamental frequency and amplitude;

- Breathy: the voice is produced with the uncompleted glottis closure that creates a audible turbulent noise;

- Asthenic: general fatigue due to insufficient muscles strain related to low voice intensity and lack of high frequency harmonics;

- Strained: iperfunctional phonetic state evaluation characterizing by noise, harmonic in the high frequency range and high fundamental frequency.

For each parameters is assigned a quote from 0 (normal) to 4 (severe).
VHI is an auto-evaluation scale, it is a standardized 30-points questionnaire divided into 3 subsection including functional, emotional and physical voice disorders. Patients have to assign a score from 0 (never) to 5 (always) to each statement, the maximum score is 120 (worst situation).

Chapter 2

Materials and Methods

In 1945 the war was just ended and there were a great amount of people suffering throughout the country. The Don Gnocchi's project was to help people in need, starting with the most needy such as war orphans and mutilated children. In the following years Don Gnocchi Foundation expanded to include more and more patients to give attention to all forms of disability. Today, Don Gnocchi Foundation includes 5700 operators with more than 3700 beds for a total of 28 Centres and around 30 territorial outpatient clinics spread over 9 Italian regions.
This thesis work is conducted in collaboration with the speech therapy and rehabilitation department of the Don Gnocchi hospital in Milan, which treats patients with multiple sclerosis (MS). The study focused primarily on comparing parameters of healthy subjects and parameters of subjects with multiple sclerosis. The aim is to find the most significant features for the distinction between pathological and healthy subjects and the identification of feature values representative of the two categories. The second purpose of this study is to prove the best effectiveness of LSVT-LOUD technique in voice therapy for PwMS in comparison with the standard technique already in use. The last analysis concerns the comparison between the results of the extraction algorithm and the experts' perceptual evaluations (GIRBAS scale) to check the consistency of the results in terms of improvement or deterioration of vocal performance. For this reason, all patients participating in the study were divided into three groups of therapy; a first group was treated with LSVT-LOUD therapy, a second group was treated with standard therapy and a

third group was not treated. The therapists recorded the patients before and after therapy while performing 3 repetitions of the vowel /a/ and free speech lasting approximately 1 minute.

2.1 Data-set

The data set provided by Don Carlo Gnocchi includes voice recordings of 69 multiple sclerosis (MS) patients; 23 were treated with the LSVT-LOUD therapy (labeled as LSVT), 20 were treated with the standard therapy (labeled as ACTIVE) and 26 were not treated at all (labeled as WAITING). For each patient there is a set of recordings acquired before the start of therapy (time T0) and a set acquired at the end of therapy (time T1). The set includes 3 repetitions of the vowel /a/ and free speech lasting approximately 1 minute. Expert perceptual ratings (GIRBAS scale) and patient self-assessment (VHI scale) associated with each recording were also provided. General information was also provided for each patient under analysis such as: age, gender, year of disease onset and other parameters related to the stage and type of disease. GIRBAS and the VHI scores relating to the two observation times were also provided.
About the comparison between MS patients and healthy subjects, a data-set available at the Electronics and Telecommunications Department of Politecnico di Torino have been processed in order to obtain reference data. The data-set includes 3 repetitions of the vowel /a/, free speech and the reading of a phonetically balanced text for each of the 57 involved healthy subjects.

2.2 Signal Processing

2.2.1 Manual Cleaning

For this study, it was decided to take the 3 repetitions of the vowel /a/ and free speech into analysis so that patients' vocal abilities could be fully assessed. Unfortunately, due to logistical and organizational problems some records are missing so it was necessary to create subgroups of analysis so as to make the best use of the provided material. In addition, some recordings were preminarly

discarded because of saturation problems or issues related to the quality of the microphone and acquisition system. After this first selection, the recordings were analyzed by means of the software Audacity, which allowed to remove signal sections that are characterised by instrumental artefacts or external noise. During manual removal, care was taken to not cut sections of the recordings during phonation: the cut should begin and end only during silences, thus avoiding the truncation of the vocal signal. Specifically, the encountered problems are:

- Saturation: the signal exceeds the measuring range of the instrument acquisition system all out-of-range information is lost. In this case the signal is discarded in its entirety. The problem can only be solved upstream by configuring the instrument correctly before acquisition.

- Instrumental artifact: signal sections with non-zero mean due to instrumental transient; in this case a manual removal of the corrupted sections is carried out without problems, since this problem always occurred at the beginning of the recording.

- Low frequency artefact: these artifacts occur as signals with different frequency characteristics from the speech signal, in which case the artifact is not deletable since it is internal to the signal but falls in a different frequency range from that of the speech signal (thus not affecting frequency analysis but could be a problem for amplitude analysis).

- High frequency artefact: these are impulsive disturbances that fall, as before, into ranges not of interest and do not affect measurements in the frequency domain.

Some visual examples of the above problems are shown in Fig: 2.1. At the end of the manual cleaning, two main subsets were created, one for the analysis of the 3 repetitions of the vowel /a/ and a second for the analysis of speech. In both cases, the recordings made at observation times T_0 and T_1 were considered since they are most representative of the therapy effects. Tables 3.1 and 3.2 show the two final subsets of patients, the class of therapy they belong to and their numerosity.

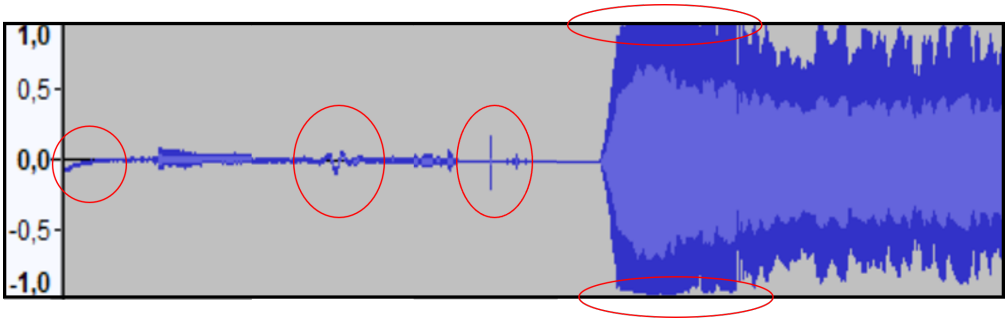

Figure 2.1: Problems encountered during manual pre-processing: a) instrumental artifacts b) low frequency artifact c) high frequency artifact d) saturation

Vowel /a/ Data-set TOTAL OF 44 PATIENTS	
Patient therapy	**Numerosity**
LSVT	14
ACTIVE	15
WAITING	15

Table 2.1: Data-set for the analysis of the vowel /a/, patients are divided in 3 therapy classes with the corresponding numerosity.

Free speech Data-set TOTAL OF 38 PATIENTS	
Patient therapy	**Numerosity**
LSVT	8
ACTIVE	16
WAITING	14

Table 2.2: Data-set for the analysis of the free speech, patients are divided in 3 therapy classes with the corresponding numerosity.

2.2.2 Pre-processing

The pre-processing of free speech and the repetition of 3 /a/ was performed in parallel using two separate scripts in Matlab R2020a environment, but the steps are equivalent. First, the signals are re-sampled at two different sampling frequencies for the two cases under analysis: in the case of the 3 repetitions of the vowel /a/ at 44100 Hz, in the case of the free speech at 22050 Hz. Then the mean value of the entire signal is removed, after verifying that it was less than 10% of the RMS value. Each signal is then normalised with respect to its maximum in order to have comparable signals. For this preliminary analysis, signals are observed using 46 ms frames; for free speech, 1024-sample windows and a sampling frequency of 22050 Hz are used (so 1024/22050 Hz = 46 ms), while for the vowel /a/, 2048-sample windows and a sampling frequency of 44100 Hz are used (analogously 2048/44100 Hz = 46 ms). At this point, the signal is divided between silence and

non-silence frames, for which a fixed threshold equal to half the RMS value of the entire signal is used: if the RMS value of the frame does not exceed half the RMS value of the entire signal, the frame is considered a silence frame. A second check is made on the frame noise, if the HNR value (the calculation will be explained in detail below) of the frame is greater than zero, i.e. the harmonic power is greater than the noise power, the frame is considered harmonic. A third check is performed on frequency jumps: frames that differ by more than half an octave are not allowed. Valid frames are saved while retaining information about the frames that were discarded from the analysis (discarded because they are silent or because they are not harmonic) so as to have a general idea of the quality of the original recording. Once the pre-processing is finished, parameter extraction is continued on the valid frames only.

2.2.3 Features Extraction

Feature extraction is performed only on the signal blocks selected during preprocessing. As mentioned in the previous paragraph two different scripts were used for feature extraction, in fact in the case of the free speech 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. The speudo-periods were calculated using the autocorrelation of the signal, which, according to theory, reaches a local maximum (the absolute maximum is at the zero shift, which corresponds to the signal's power) at the shift value equal to the signal's period, i.e. where the signal repeats more or less the same. The fundamental frequency and HNR, as well as the speudo periods are also calculated using the autocorrelation method. In the following section, its steps are briefly explained. Four acoustic parameters were used to analyse and compare the vocal performances of the patients: RMS, HNR, f_0 and CPPS. The parameters were observed through their relative statistical distribution, considering for each of them the following statistics: mean, median, mode as central trend; standard deviation, range, 5° percentile, 95° percentile as variability measure; skewness and kurtosis as shape factor. In the case of the vowel /a/, 9 stability parameters in period and amplitude are also calculated. For the Matlab implementation of the parameters, reference was made to the software instruction manual MDVP, Model

5105 [8].

2.2.4 Acoustic Parameters

RMS: Root Mean Square

The RMS value is used in zero mean alternating signals as a power indicator. The function already implemented in Matlab "rms" was used for the calculation. Analytically:

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

where x_n are the signal samples. The RMS value of each frame is calculated before the extraction of the other acoustic parameters as a discriminator for the silent frames as explained above. The RMS value of the non-silent frames is then saved in an array.

HNR: Harmonic to Noise Ratio and f_0 Fundamental Frequency

Average ratio of the harmonic spectral energy. It provides a general evaluation of the presence of noise in the signal. For the calculation of the HNR and the fundamental frequency f_0, the autocorrelation method was used; to understand the various steps, it is necessary to recall the definition and properties of autocorrelation of a generic signal.

$y(t)$ is a generic steady-state signal in the time domain, i.e. a signal whose probability distribution does not change when shifted in time. A generic speech signal consists of a periodic part, which therefore repeats itself after a period of time equal to the duration of its period, and a non-periodic part due to noise. the autocorrelation of a signal is defined as the product of the signal by itself shifted by a quantity m varying between zero and the length of the signal. The autocorrelation (equation 2.2) is defined in the time-domain, consequently it has an absolute maximum at zero (completely overlapping signal) representing the signal power and is zero where the signals do not overlap. For signals that have a periodic component as in this case, the autocorrelation also has a local maximum for shift value m equal to the period of the signal, i.e. where the signal reoccurs more or less the same. So by identifying the point of local maximum, it is possible to obtain the

period of the signal and consequently also its fundamental frequency f_0. To speed up computational time, the search is carried out in the period ranges corresponding to the frequency ranges of interest; 75 Hz to 300 Hz for males, 100 Hz to 400 Hz for females.

$$AC = \int y(t) * y(t+m)\, dt \tag{2.2}$$

once the period T is found, the parameter HNR is calculated as:

$$HNR\ [\text{dB}] = 10 \log_{10} |\frac{AC(T)}{AC(0) - AC(T)}| \tag{2.3}$$

where:

- $AC(T)$: harmonic energy of the signal, calculated as the autocorrelation of the signal at the shift point equal to the period.
- $AC(0)$: total signal energy, calcolated as the autocorrelation in the zero shift.
- $AC(0) - AC(T)$: portion of energy related to the unharmonic part of the signal, calculated as the difference of the previous energies.

HNR is evaluated in dB through the use of the logarithmic operation: when it is greater than zero, it means that the periodic signal power exceeds the non-periodic signal power. The HNR value is calculated on non-silent frames to assess their harmonicity: if the value exceeds 0 dB the frame is accepted and the HNR value saved in an array and extraction is continued; if it does not exceed zero it is considered a non-harmonic frame.

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. "Cepstrum" is the anagram of the word "spectrum", it is defined as the power spectrum of the logarithm of the signal power spectrum ([9], [10]). The power spectrum is defined as the square of the Fourier transform, so applying the squared transform to the logarithm of the squared transform of the signal gives the cepstrum:

$$C_p(\tau) = \mathcal{F}\{\log |\mathcal{F}[x(t)]|^2\}^2 \tag{2.4}$$

where: $x(t)$ is the vocal signal, $\mathcal{F}$ is the Fourier transform, $|\mathcal{F}[x(t)]|^2$ is the signal power spectrum and τ is is called "quefrency" (namely the anagram of "frequency") in the cepstrum domain, being neither a time nor a frequency. The use of the cepstrum is extremely advantageous in the analysis of multi-component sounds, as in the case of vocal analysis. By switching to the quefrency domain, the complex signal is represented broken down into its simple and now easily recognisable components. For example, a generic vocal signal $y(t)$ is composed of two components, one $s(t)$ given by the vocal source (vocal cords) and a second $h(t)$ given by the resonant tract (which filters the first signal) in convolution with each other:

$$y(t) = s(t) * h(t) \tag{2.5}$$

By applying the Fourier transform to the equation (2.5), the convolution is transformed into a multiplication, obtaining:

$$Y(f) = S(f) \cdot H(f) \tag{2.6}$$

Applying the logarithmic function to (2.6) gives:

$$\log Y(f) = \log S(f) + \log H(f) \tag{2.7}$$

In this way, it is possible to obtain the equation (2.7) that has the two components in sum and no longer in multiplication. Finally, reapplying the Fourier transform results:

$$\mathcal{F}[\log Y(f)] = \mathcal{F}[\log S(f)] + \mathcal{F}[\log H(f)] \tag{2.8}$$

the graphical representation (fig.2.2) of the cepstrum, which shows on the abscissae the quefrency in [ms] and on the ordinates the amplitude in [dB], makes it possible to identify a peak in amplitude localised around the fundamental period of the vocal signal $y(t)$. The CPPS parameter is obtained by measuring the difference between the amplitude of the cepstrum peak, on which smoothing is first performed, and the value of the cepstrum regression line at the peak point. Smoothing was introduced a posteriori and showed a clear improvement in results.
For the implementation of the CPPS factor, it was necessary to proceed in a different way to the other parameters, as many more values had to be calculated

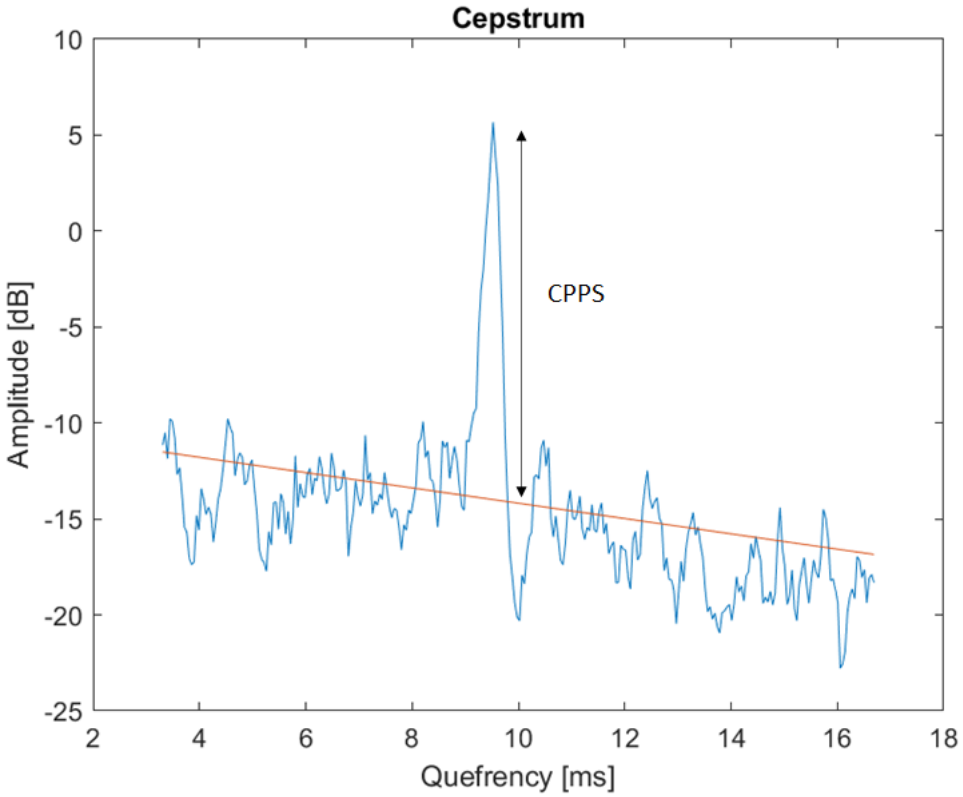

Figure 2.2: In blue the cepstrum extracted from a patient speech recording, in red the related regression line, the quefrency in which the peak falls corresponds to the fundamental period of the signal.

in order to obtain satisfactory results. To do this, the CPPS was calculated every 44 samples, but using signal hamming windows of 1024 samples, this created an indirect overlap factor of $(1024 - 44)/1024 = 96\ \%$. Once the cepstrum is smoothed over 7 frames, the CPPS can be calculated as the difference between the regression line on which the cepstrum is smoothed and its maximum value. To speed up the implementation and to avoid errors, the search for the maximum is only done in the range of interest, which falls between 3.3 ms and 16.7 ms.

2.2.5 Recording Parameters

Three pre-processing output parameters are also saved, representing the percentage of frames rejected in the two pre-evaluation steps prior to feature extraction; they are useful for the interpretation of the rest of the parameters. In particular, the

stability parameters that follow are greatly influenced by the length of the signal under analysis; in fact, being the result of an averaging operation, the longer the signal, the more stable parameters will result. Below are the formulas used:

- $\frac{V}{S}$: ratio of non-silent frames, which include both harmonic and non-harmonic frames, by silent frames (considered only for the free speech);
- $\frac{har}{har+unhar}$: ratio of harmonic frames by non-silent frames;
- Length: number of valid frames after pre-processing.

2.2.6 Stability Parameters

The parameters used for the analysis of the 3 repetitions of the vowel /a/ and free speech are the same, with the exception of 9 additional parameters of amplitude and period stability for the vowel /a/ analysis. In fact these parameters are useful to get an idea of how much the patient is able to keep the tone and degree of the voice constant. The 9 parameters and their corresponding definitions are given below:

1. **Jitta** [μs]: Absolute Jitter, an evaluation of the period-to-period variability of the pitch period. Voice break areas are excluded.

$$Jitta[\mu s] = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}| \tag{2.9}$$

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

2. **Jitt** [%]: Jitter Percent, relative evaluation of the period-to-period variability.

$$Jitt[\%] = 100 \cdot \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |T_0^{(i)} - T_0^{(i+1)}|}{\frac{1}{N} \sum_{i=1}^{N} T_0^{(i)}} \tag{2.10}$$

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

Both Jita and Jitt represent assessments of the same type of pitch disruption. Period irregularity can be associated with the inability of the vocal cords to sustain a periodic vibration for a given period. Usually these types of variations are random. They are typically associated with hoarse voices. Jita is an absolute measure dependent on the average fundamental frequency of the voice. For this reason, normative values from the Jita for men and women differ considerably. Higher pitch is associated with lower Jita and vice versa, which makes Jita difficult to compare. Jitt, on the other hand, is a relative measure and the influence of the subject's average fundamental frequency is greatly reduced.

3. **RAP** [%]: Relative Average Perturbation, Relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with smoothing factor of 3 periods.

$$RAP[\%] = 100 \cdot \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{T_0^{(i-1)}+T_0^{(i)}+T_0^{(i+1)}}{3} - T_0^{(i)}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_0^{(i)}} \tag{2.11}$$

where: $T_0^{(i)}, i = 1,2...N$ extracted pitch period data,
N: number of extracted pitch periods. It is similar to the Jitt but with a smoothing that reduces the sensitivity of RAP to pitch extraction errors. However, it is less sensitive to very short-term variations. Hoarse and/or breathy voices may have a higher RAP.

4. **PPQ** [%] : Pitch Period Perturbation Quotient, Relative evaluation of the period-to-period variability of the pitch within the analyzed voice sample with a smoothing factor of 5 periods.

$$PPQ[\%] = 100 \cdot \frac{\frac{1}{N-4}\sum_{i=1}^{N-4}\left|\frac{1}{5}\sum_{r=0}^{4} T_0^{(i+r)} - T_0^{(i+2)}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_0^{(i)}} \tag{2.12}$$

where: $T_0^{(i)}, i = 1,2...N$ extracted pitch period data,
N: number of extracted pitch periods. PPQ is a parameter very similar to

RAP only it has smoothing over 5 periods instead of 3; thus the effect of smoothing is more predominant.

5. vf_0 [%]: Coefficient of Fundamental Frequency Variation, Relative standard deviation of the fundamental frequency. It reflects, in general, the variation of f_0 (short to long-term) within the analyzed voice sample. vf_O is computed as the ratio of the standard deviation of the extracted period-to-period fundamental frequency data by the average fundamental frequency as:

$$vF_0[\%] = 100 \cdot \frac{\sigma}{f_0} = 100 \cdot \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N}\sum_{j=1}^{N} f_0^{(j)} - f_0^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N} f_0^{(i)}} \tag{2.13}$$

where: $f_0 = \frac{1}{N}\sum_{i=1}^{N} f_0^{(i)}$ and $f_0^{(i)} = \frac{1}{T_0^{(i)}}$ period to period fundamental frequency values, where: $T_0^{(i)}, i = 1,2...N$ extracted pitch period data, N: number of extracted pitch periods. vf_O reveals changes in the fundamental frequency. The vFo value increases regardless of the type of pitch variation. Random or regular variations, short or long term, increase the vFo value. These variations can be frequency tremors or non-periodic variations, or even simply an increase or decrease in pitch.

6. **Shim** [%]: Shimmer Percent, Relative evaluation of the period-to-period (very short term) variability of the peak-to-peak amplitude within the analyzed voice sample.

$$Shim[\%] = 100 \cdot \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A^{(i)} - A^{(i+1)}|}{\frac{1}{N}\sum_{i=1}^{N} A^{(i)}} \tag{2.14}$$

where: $A(i), i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. Cycle-to-cycle amplitude irregularity may be associated with the inability of the strings to sustain a periodic vibration for a given period and the presence of turbulent noise. This type of random irregularity is typically associated with hoarse and breathy voices.

7. **ShdB** [dB]: Shimmer in dB - Evaluation in dB of the period-to-period (very

short-term) variability of the peak-to-peak amplitude within the analyzed voice sample.

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

where: $A(i), i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. Both Shim and ShbB are relative evaluations of the same type of amplitude perturbation, but use two different units, percent and dB.

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

$$APQ[\%] = 100 \cdot \frac{\frac{1}{N-10}\sum_{i=1}^{N-10}|\frac{1}{11}\sum_{r=0}^{10}A^{(i+r)} - A^{(i+5)}|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.16}$$

where: $A(i), i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. APQ is a parameter very similar to shimmer but with a smoothing factor of 11. Although smoothing reduces the sensitivity of APQ to period-to-period amplitude variations, APQ still describes short-term amplitude perturbations of the voice very well. Wheezy and hoarse voices usually have a higher APQ.

9. **vAm** [%]: Coefficient of Amplitude Variation, Relative standard deviation of the peak-to-peak amplitude. It reflects in general the peak-to-peak amplitude variations (short to long-term) within the analyzed voice sample.

$$vAm[\%] = 100 \cdot \frac{\sigma}{A_0} = 100 \cdot \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\frac{1}{N}\sum_{j=1}^{N}A^{(j)} - A^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.17}$$

where: $A(i), i = 1, 2...N$ extracted peak to peak amplitude data, N: number of extracted impulses. vAm reveals cycle-to-cycle amplitude variations of the voice. Any type of variation, short or long term, regular or random, increases

the value of vAm.

As a result, matrices are obtained which show the time observations of the various patients under analysis on the rows and the extracted parameters on the columns. In the case of the free speech on the columns, 39 features are reported, whereas for the case of the 3 /a/ repetitions there are a total of 47 features.

2.3 Data cleaning (vowel /a/)

Following the extraction, the parameters were checked to verify their quality and detect any anomalous values. The most variable and far from the respective standard values turn out to be those for the stability parameters. The stability parameters (calculated only in the case of the vowel /a/), in particular, assume higher values where the recording under analysis is longer, i.e. where there are more signal windows under analysis. The number of windows under analysis is reported on the last column of the parameter matrix obtained from the extraction; this number is obtained by subtracting the number of silence windows and the number of non-harmonic windows from the total number of signal windows. Furthermore, the stability parameters are calculated on consecutive frames, but if some frames internal to the signal are removed during preprocessing, frames that are not actually adjacent are considered as such and this leads to errors in the calculation of the parameters. In order to quantify the effect of this undesired phenomenon, the frames analysed by the pre-processing are represented in a graph by assigning a value of 1 to harmonic frames, a value of 0 to silent frames and a value of 2 to non-harmonic frames, two examples are given in Fig:2.3. Ideally, one would expect to observe the presence of silent frames at the beginning and end of the recording, but in the middle of the vocal reproduction, only harmonic frames should be present; what is observed in many cases, is the presence of numerous non-harmonic frames within the recording and sometimes even silence frames. This situation should normally not occur because every audio reports sustained phonation of the vowel /a/ (harmonic sound, produced by the periodic vibration of the vocal cords). Obviously, these are pathological voices, so the patient may lose his voice at times (silent frames) or the sound may not be clear (non-harmonic frames). On the other hand the quality of the starting audio was poor (a lot of sub-frame noise) and many

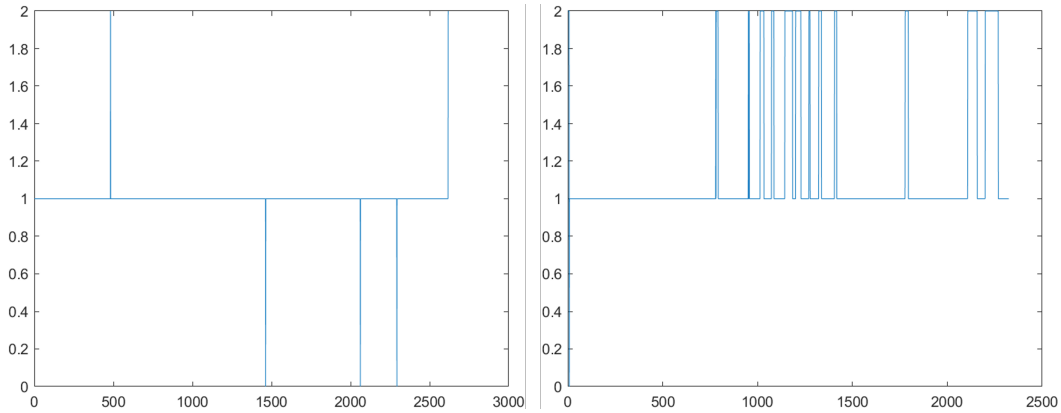

Figure 2.3: Two examples relating to the recording of the vowel /a/ sustained, on the x's are the total frames on the y's are associated the value: 1) harmonic frame, 2) non-harmonic and 0) silence.

frames within the signal are considered non-harmonic and eliminated, creating holes that lead to a consistent worsening of the stability parameters; moreover, the recording is shorter and also in this case the stability parameters are worse. It was therefore decided to eliminate the most problematic recordings where the whole recording was discontinuous, and in cases where a more harmonious and continuous inner tract was observed to cut the beginning and end parts. After this cleaning, the recordings were processed again and the parameters were save. By having 3 values for each observation time, it was possible to eliminate unsuitable data (one or two repetitions out of 3 total) and save suitable ones, without excluding the patient from the analysis.

2.4 Pathological vs. Healthy subjects

To get an idea of the pathological vocal situation of the patients, a comparison with healthy subjects was made. The Turin Polytechnic provided recordings of 57 healthy subjects, for each of whom there were three repetitions of the vowel /a/, a free speech and the reading of a phonetically balanced passage. The same parameter extraction algorithm was used for both the vowel /a/ and the free speech (in this case, reading was also analysed in the same way as speech). The parameters extracted from the 3 repetitions of the vowel /a/ were averaged and

standard deviations were calculated. The data were then observed by means of point cloud graphs in order to observe a group trend and to understand which features differed most between healthy and unhealthy subjects. Referring to the fig:(2.4), pathological patients are represented in colours ranging from yellow to red, with lighter colours indicating a low G grade and thus better conditions, while a colour closer to red indicates a high G grade and thus a more severe pathological situation; the colour blue is used to represent healthy patients. On the part of the graph where the pathological patients are represented, there are red vertical lines that divide the pathological patients into 6 subgroups, respectively: pre-therapy and post-therapy LSVT-LOUD patients, pre-therapy and post-therapy STANDARD patients and patients not treated with any therapy before and after few months.

From such graphs, a cut-off value could be identified for the most significant

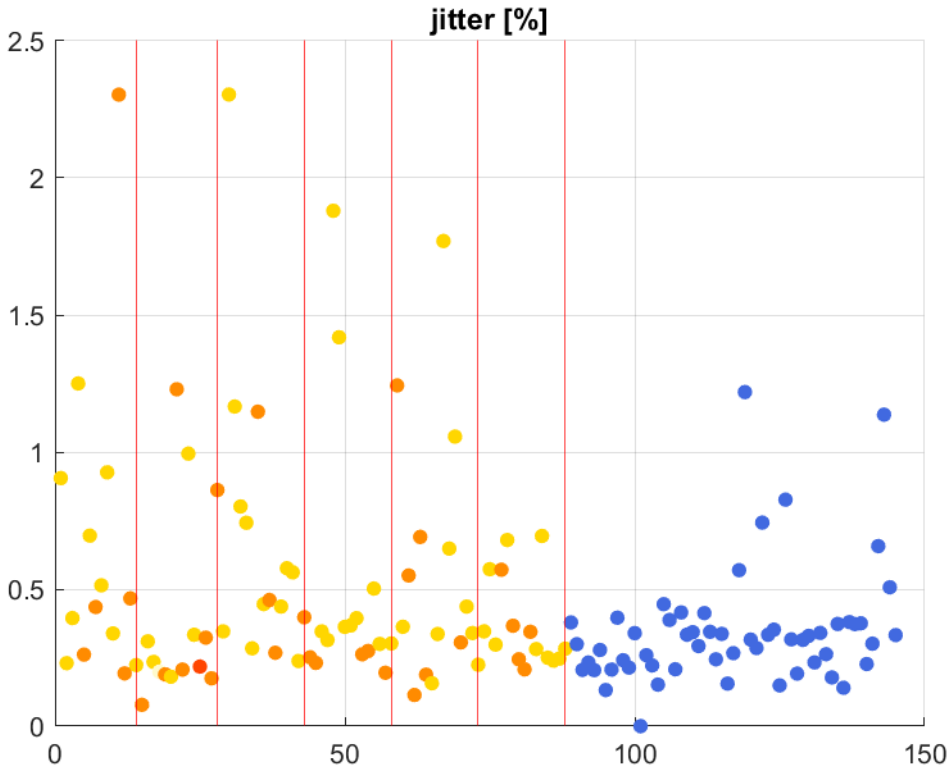

Figure 2.4: Absolute value of jitter % of pathological patients (yellow to red as the severity of vocal ability increases) and healthy subjects in blue

features to differentiate healthy from sick patients. From this idea, four groups

were created:

1. All pathological patients before therapy;

2. Pathological patients after therapy (both LSVT-LOUD and standard);

3. Untreated pathological patients;

4. Healthy patients.

On these four groups, the mean and the standard deviation of the mean were calculated and the values compared by means of bar graphs, fig:2.5. With these types of graphs, it is possible to quantify the improvement in the vocal performance of pathological patients and to get an idea of how close this improvement is to a non-pathological situation.

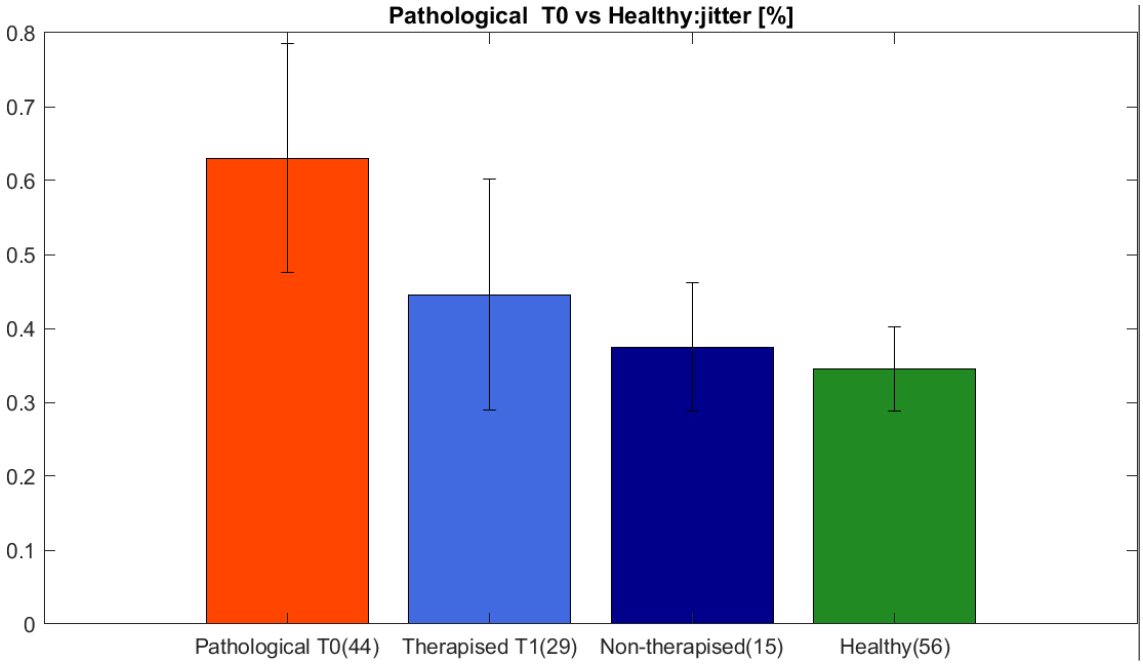

Figure 2.5: Average value and dispersion ($\pm 2\sigma$) of jitter with reference to the 4 categories under analysis.

2.5 Feature Selection

2.5.1 Data Observation

Before proceeding with feature extraction, a preliminary observation of the data was carried out. For the analysis of /a/, 3 repetitions were available for each time observation, so the mean value and standard deviation were calculated, which will be used from this point onwards for the analysis. The graph (Fig:2.6) shows the average values of the 3 repetitions and the standard deviations through the error bars of amplitude $\pm\sigma$ are also shown for the parameter HNR mean. The same type of graphical observation was made with the parameters extracted from the free speech analysis, but in this case there was just one recording for each observation time, so the standard deviation was not represented in the graphs. This first visual observation is intended to help the observer compare the patient's situation before and after therapy, highlighting any improvements or worsening. At this point, the variations between the two temporal observations T_1 and T_0 were calculated by difference; in the case of the free speech the parameter matrices of the two observations were simply subtracted according to the formula T_1 - T_0; in the case of the vowel /a/ the mean values of the 3 repetitions, previously calculated, were subtracted in a similar way, and the standard deviations:

$$\sigma = \sqrt{\sigma_{T_0}^2 + \sigma_{T_1}^2} \tag{2.18}$$

obtained by quadratic summation from the standard deviations of the 2 temporal observations were associated to those delta values. In order to better understand what has just been illustrated, an example is given that is related to the parameter CPPS mean for free speech (Fig:2.8) and the parameter HNR mean for the vowel /a/ (Fig:2.7).

In Fig:2.8 and Fig:2.7, where the parameters CPPS mean and HNR mean are represented, positive values indicate improvements in the patient voice (patient number shown on the abscissa) as a result of the rehabilitation.
At this point, it was decided to observe the average class values (LSVT, ACTIVE and WAITING) to note any differences in results in relation to the patient rehabilitation. In the case of free speech, the delta values for each different feature were simply

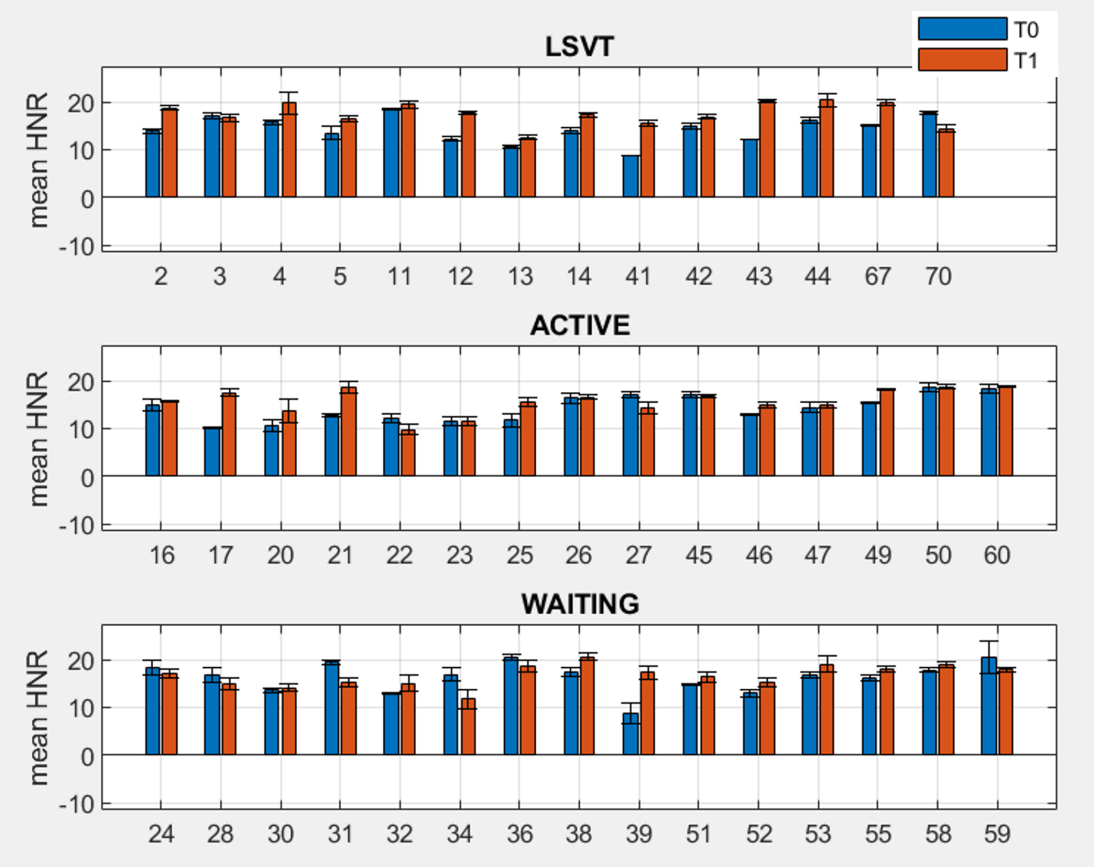

Figure 2.6: Mean value over the 3 repetitions of the vowel /a/ of the mean HNR over the two time observations T_0 and T_1 for the patients in the analysis divided by class.

averaged across the three classes and the standard deviation of the mean was also calculated. For the vowel /a/ the delta values were averaged across the three classes in a similar way, whereas two contributions had to be taken into account to calculate the overall dispersion:

- Intra-class contribution u_1, which is related to the dispersion of each subject with respect to the class mean-value;

- Intra-subject contribution u_2, which takes into account the standard deviation of each subject that belongs to the class.

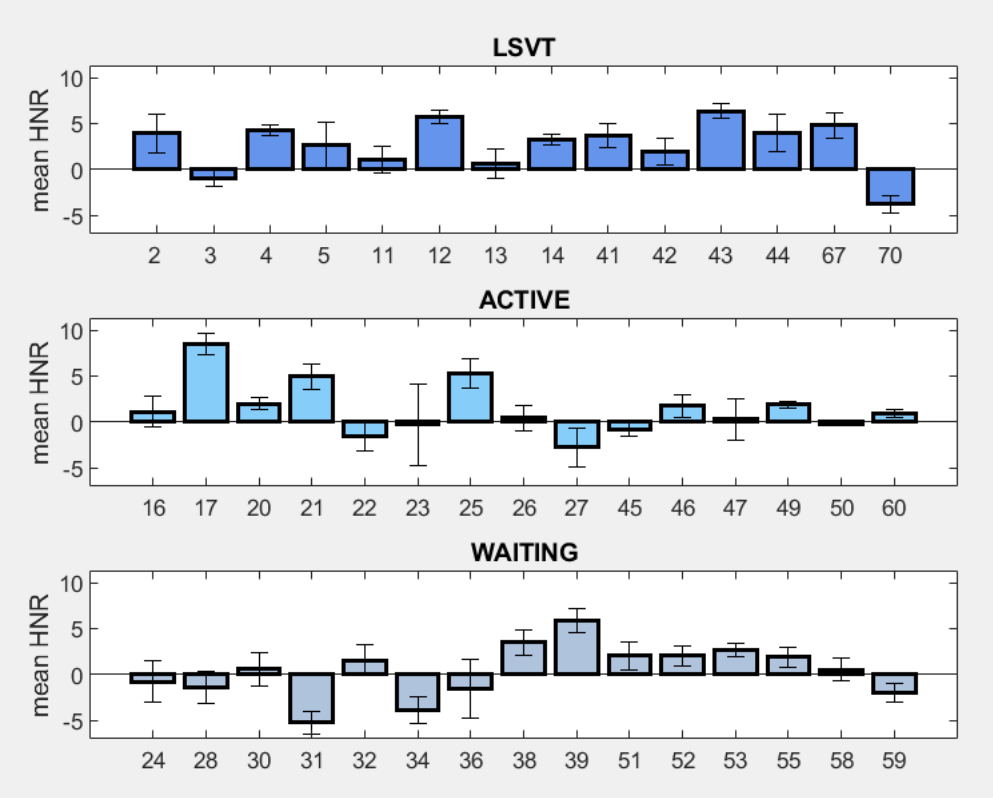

Figure 2.7: Vowel /a/: Mean HNR delta values ($T_1 - T_0$) and relative standard deviations of the subjects divided into the three classes.

$$u_1 = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\delta_i - \bar{\delta})^2} \qquad u_2 = \sqrt{\frac{1}{N^2}\sum_{i=1}^{N}\sigma_i^2}$$

Where N is the number of elements belonging to each class. These two contributions were then summed quadratically to obtain the total dispersion:

$$u_{tot} = \sqrt{u_1^2 + u_2^2} \tag{2.19}$$

To better understand what has just been described, visual examples are shown in Fig:2.10 and Fig:2.9 that refer to th parameters CPPS mean for the vowel /a/ and HNR mean for the free speech, respectively.

The observation of these graphs, relating to each extracted feature, aims at the

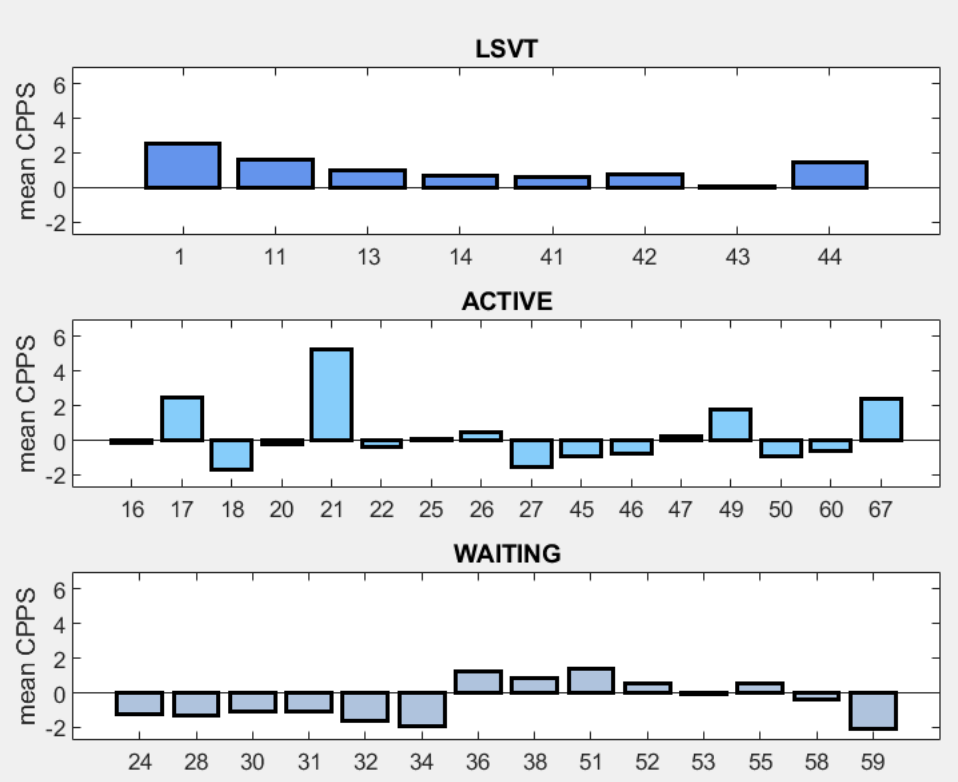

Figure 2.8: Free speech: Mean CPPS delta values ($T_1 - T_0$) of the subjects divided into the three classes.

identification of specific features allowing a division between classes. To this end, the average trends with their associated dispersion bands were observed. Indeed, if features with dispersion bands that do not intersect between the three classes were identified, these could be selected for the classification of patients into the three classes.

In the case of the vowel /a/, delta intervals at $\pm 2\sigma$ level do not overlap for the parameters HNR mean and median, while for free speech the delta intervals are overlapped for all the parameters. Moreover this type of Features Selection allows features to be considered individually and not in their combinations. This way, the discriminatory power of the features is not maximised. For this reason, a combinatorial algorithm, based on Logistic Regression was used to perform Features Selection. The algorithm selects all the possible features and all the possible combination of 2, 3 and 4 features.

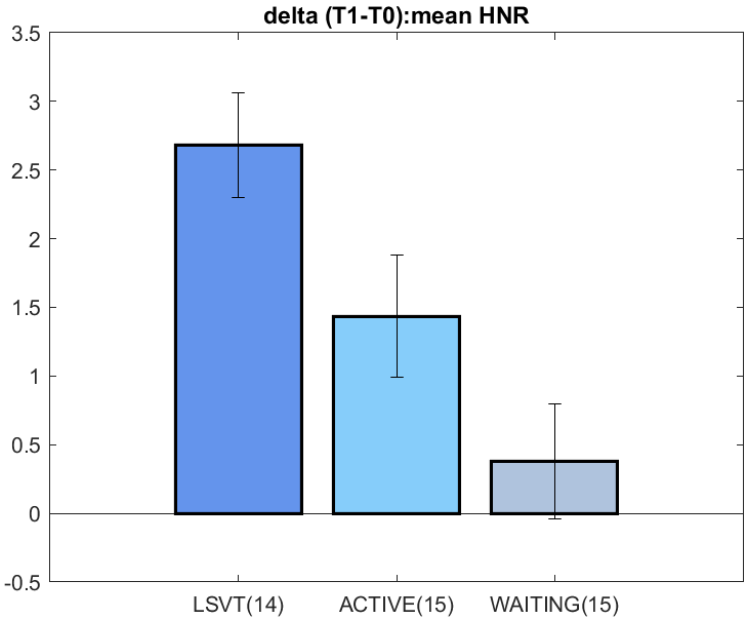

Figure 2.9: Vowel /a/: Average feature values with relative dispersion considering the range $\pm\sigma$, on the abscissas are the classes with the total number of patients belonging to them.

2.5.2 Logistic Regression

Logistic Regression (LR) is a non-linear regression model used when the dependent variable is a binary value; in fact, unlike linear regression, which returns values belonging to the set of real numbers $\Re$, logistic regression uses a probability that by definition is limited in the interval [0,1]. Logistic regression can be used to classify observations, based on their features, into the two categories. LR is a supervised classification algorithm widely used in machine learning; these are algorithms that train using a complete data set of features with associated class. The objective of the model is to establish the probability with which an observation x (independent variable) can generate one or the other value of the dependent variable y (0 or 1). The logistic, or logit, model associates a logarithmic probability function (2.20) with a linear combination of independent variables X_i and regression coefficients β_i, where β_0 is the intercept and $i = 1...N$ with N: number of observation.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_N x_N = f(x) \tag{2.20}$$

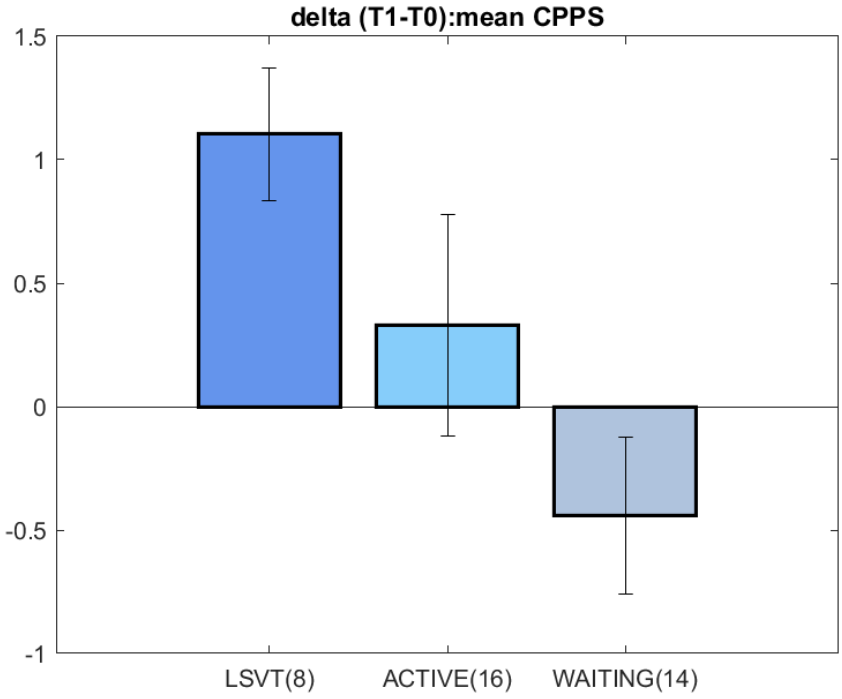

Figure 2.10: Free speech: Average feature values with relative dispersion considering the range $\pm\sigma$, on the abscissas are the classes with the total number of patients belonging to them.

The probability returned by the algorithm is defined by the sigmoid function Eq.2.21 derived from Eq.2.20, this is a continuous function defined between 0 and 1:

$$p(x) = \frac{1}{1 + e^{-f(x)}} \tag{2.21}$$

The objective of the algorithm is to minimise classification errors [11], by finding a sigmoid function $p(x)$, defined by the regression coefficients $\beta_0, \beta_1 ... \beta_N$, that can minimise the distance between it and the actual responses Y_i. The process of calculating the best coefficients or weights is called model training. In order to obtain the best weights, the maximum likelihood is estimated by maximising the log-likelihood function, LLF (equation 2.22); this can be done with various deterministic or stochastic mathematical methods such as least squares difference,

gradient descent and Newton's method:

$$LLF = \sum_{i}^{N} (y_i \ln(p(x_i)) + (1 - y_i) \ln(1 - p(x_i))) \tag{2.22}$$

The relationship between $p(x)$ and $f(x)$ in Eq. 2.20 implies that $p(x) = 0.5$ when $f(x) = 0$ and in this anointing the threshold is imposed: the expected output is 1 if $f(x) > 0$ ($p(x) > 0.5$) and 0 otherwise ($p(x) <= 0.5$). The threshold need not be 0.5, but it usually is, otherwise a lower or higher value can be defined if it is more convenient for the specific situation. For better understanding, the simplest case of logistic regression, single-variable logistic regression, is taken into account. In this model only one feature is considered and thus the logit function results in:

$$f(x) = \beta_0 + \beta_1 x \tag{2.23}$$

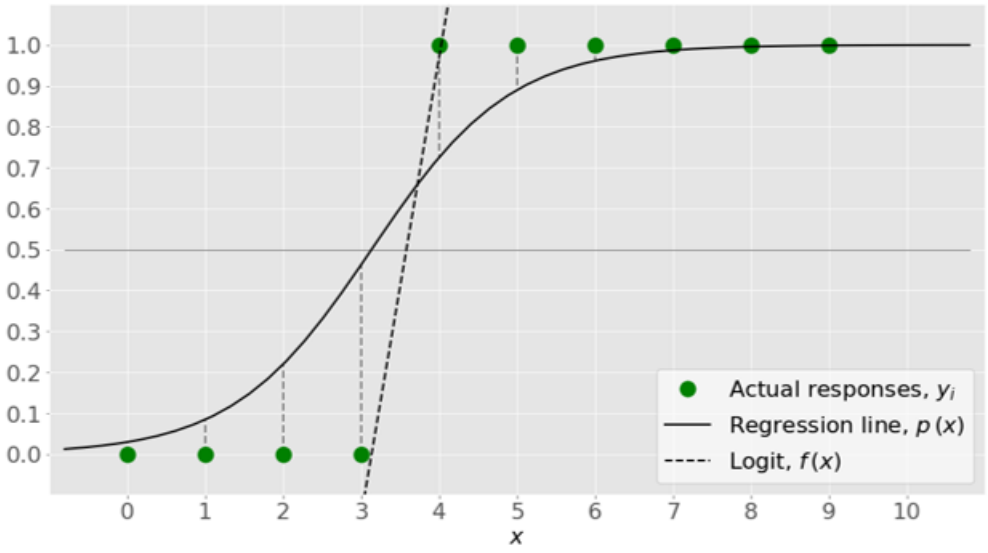

Figure 2.11: Single-variate Logistic Regression model.

The Fig:2.11 shows a visual example of what has just been illustrated; binary true answers are shown in green, the sigmoid probability function $p(x)$ is shown as a continuous black line and the logit function $f(x)$ as a dotted line.

Once the prediction results have been obtained, it is possible to evaluate performance by comparing the results predicted by the algorithm with the true results. The comparison of the results obtained from the algorithm with the actual results yields the Confusion Matrix (CM) (see Fig:2.12), in which 4 result categories are compared, the totality of which must return the total number of observations. In fact, if we consider the binary encoding 0 - 1 as Negative - Positive respectively, the algorithm can act in 4 different ways with the binary data: the input data is Positive and the algorithm returns a correct Positive value, vice versa it could erroneously return a Negative value, similarly if the input data is Negative the algorithm could return a correct Negative data, or be wrong and return a Positive data. Consequently are defined:

- True Positives (TP): Total of correctly classified Positive values;
- True Negatives (TN): Totality of correctly classified negative values;
- False Positives (FP): Totality of incorrectly classified Negative values;
- False Negatives (FN): Totality of wrongly classified Positive values.

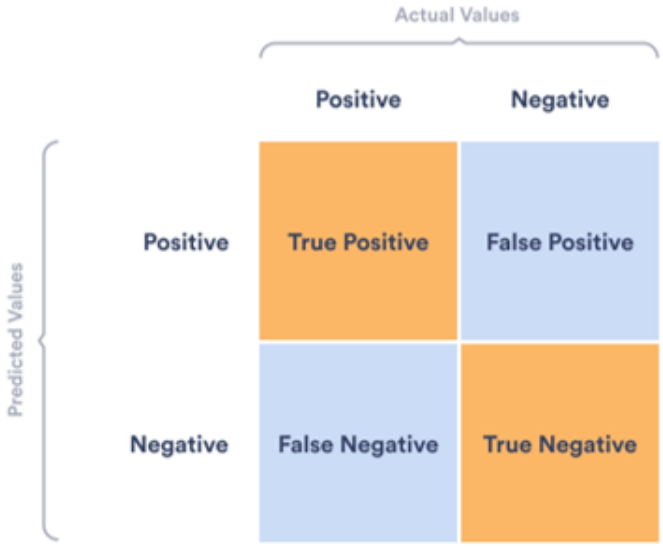

Figure 2.12: Confusion Matrix in binary classification.

From these 4 response types, metrics are defined to assess the algorithm's capabilities and performance. The most widely used metrics are:

- Accuracy [%]:

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

  Percentage of correct classified in both categories, indicates the overall goodness of classification.

- Precision [%]:

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

  Percentage of correct among all positive classifiers. It measures the algorithm's ability to classify positive samples.

- Sensitivity [%]:

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

  Sensitivity indicates the algorithm's ability to correctly classify subjects as Positive, which is very relevant if the Positive class is that of sick subjects and Negative, vice versa, is that of healthy subjects. A high sensitivity in fact corresponds to a low number of FNs, i.e. sick subjects classified as healthy, which in practice is the most dangerous situation that can occur.

- Specificity [%]:

$$SPEC = 100 \cdot \frac{TN}{TN + FP} = TNR = 1 - FPR \tag{2.27}$$

  specificity, on the other hand, is the ability to correctly classify Negative subjects. If, as before, one considers the class of Positives as that of the sick and that of Negatives as that of the healthy, high specificity translates into the ability to recognise a healthy patient as such; this is much less relevant than the ability to recognise a sick person as such; in fact, a healthy person declared sick will undergo further examinations that will disprove the illness, whereas a sick person declared healthy will not undergo further investigations with likely serious consequences.

These considerations are made in a general sense to emphasise that the capabilities

of an algorithm can often be, or are intended to be, unbalanced with respect to classes. In this study, we want to examine the different efficacies of two speech-language rehabilitation techniques, so there is no one class to favour because it is more at risk. In this case, the classification aims to distinguish the two classes of therapy as evidence and confirmation of the different impact on patients.

- Area Under Curve (AUC):

  The ROC curve is defined by plotting the true positive rate (TPR) against the false positive rate (FPR) for each possible cut-off setting (in ours case it is 0.5, but it can be moved). The true positive rate is also known as sensitivity. The false positive rate is also known as the false alarm probability and can be calculated as (1 - specificity). The ROC curve is defined between

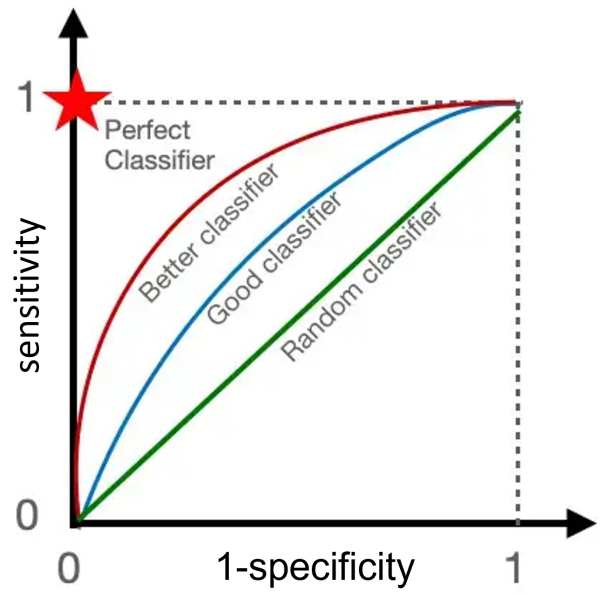

Figure 2.13: Examples of ROC curves of different classifiers [12].

  values in the interval [0,1] of abscissas and ordinates; if the curve corresponds to the diagonal, the probability of obtaining a correct result is 50 per cent, which is the worst situation and corresponds to random classification (like flipping a coin); the ideal situation of perfect classification corresponds to the co-ordinate point (0,1) (instead of the step being zero only at the zero abscissa point and 1 elsewhere); the closer the curve is to the ideal form, the

more preferable the classification, with visual reference in Fig:2.13. Sometimes, however, dubious situations arise where it is not possible to state with certainty which of the curves is preferable, thus the parameter Area Under Curve (AUC) was introduced, i.e. the integral of the curve, a value in the range [0,1], the higher the value the better the classification.

2.5.3 Feature Selection using Logistic Regression

To perform the feature selection, a classification algorithm based on logistic regression is used in the Matlab R2020a environment. As mentioned above, logistic regression is a supervised classification algorithm, which means that it requires as input elements described by a certain feature number and class. It was also mentioned that logistic regression is a binary classification algorithm; in this study, there are three classes to discriminate (LSVT, ACTIVE and WAITING), which is why three feature selections are carried out, one for each combination of classes. A matrix made up of two classes of elements at a time is given as input to the algorithm. The matrix shows on the rows the subjects belonging to the two classes and on the columns the features, the last column indicates the class. The table 2.3 shows the 3 class combinations for binary classification with their respective labels and the number of patients belonging to the classes for the analyses performed on the vowel /a/ and free speech. What is noticeable from the table 2.3 is that the

Labels	**Classes**	**Vowel /a/**	**Free speech**
0 *vs* 1	LSVT *vs* ACTIVE	14 *vs* 15	8 *vs* 14
0 *vs* 2	LSVT *vs* WAITING	14 *vs* 15	8 *vs* 14
1 *vs* 2	ACTIVE *vs* WAITING	15 *vs* 15	16 *vs* 14

Table 2.3: The 3 class combinations for binary classification with their respective labels, the two columns on the right show the class numerosity for the two analyses performed.

elements available for classification are generally few, although the class numerosity is fairly balanced, so very good results are not expected. To evaluate the best feature or combination of features for classification, the algorithm selects one or

more (up to four) features, trying out all possible combinations, and calculates the metrics described above to assess their performance. The algorithm first calculates the autocorrelation of all features using the *'corr'* function available in the R2020a Matlab environment. The function takes the matrix containing the features as input and uses Pearson's Linear Correlation Coefficient to calculate the autocorrelation Eq:2.28.

$$Corr = \frac{\sum_{i=1}^{n}(F_{a,i} - \bar{F}_a)(F_{b,i} - \bar{F}_b)}{\sqrt{\sum_{i=1}^{n}(F_{a,i} - \bar{F}_a)^2 \sum_{i=1}^{n}(F_{b,i} - \bar{F}_b)^2}} \tag{2.28}$$

Where $\bar{X}_a = \sum_{i=1}^{n}(X_{a,i})/n$ is the mean value of the a column, $\bar{X}_b = \sum_{i=1}^{n}(X_{b,i})/n$ is the mean value of the b column and n is the column length. This yields a symmetric matrix mxm (where m is the number of features) of correlation coefficients for each couple of features; values are defined between 1 and -1 (1: positive perfect correlation, 0: no correlation, -1: negative perfect correlation). A second matrix with the same dimensions is also obtained containing the p-value for the corresponding element; if the p-value is less than 0.05 the correlation is significant. The first check is performed (only in the case where the features considered are 2 or more) on correlation and p-value: if at the same time the square correlation value is between 0 and 0.5 and the corresponding p-value is less than 0.05, then the feature pair is accepted. Then this first check is performed on all possible pairs of features by selecting 1 to 4 at a time. If all possible pairs (2 features: 1 possible couple, 3 features: 3 possible couple, 4 features: 6 possible couple) of features of the specific combination satisfy the condition then the pair, triplet or quatern is considered valid and sent as input to the logistic regression model. The *'fitglm'* function available in Matlab version R2020a was used to create the model, specifying that the response variable is a binomial type and that the link function is the *'logit'* function. Once the model has been obtained, a second check is made on the extracted p-value, this time however a much higher than the previous one fixed threshold is used; in fact, it was noted that for low p-values equal to 0.05 no combination of features was selected, therefore it was decided to raise the threshold up to a value that would allow obtaining at least one triplet. The model returns for each observation a probability value between 0 and 1 which through a fixed threshold set at 0.5 is converted into a binary value 0/1 which represents the membership predicted by the model. From the comparison between the real class

and the one predicted by the model, the metrics described above are calculated and used to select the best single feature or pair or triplet or quadruple of features for classification. To make this selection, the combination of features with the highest accuracy value is taken and if there is more than one with maximum accuracy, the one with maximum AUC is selected.

Weighted LR model for the Feature Selection of /a/

The features used in the FS, both in the case of free speech and in the case of the vowel /a/, are the delta values; these were obtained by subtraction of the parameters obtained from the feature extraction of the recordings made at times T_0 and T_1 according to the formula $T_1 - T_0$. In the case of the vowel /a/, the standard deviations of the 3 repetitions were also available, from which a dispersion matrix of the same size as the delta matrix was derived, as explained above. It is therefore possible to associate each delta value with its respective dispersion value. The *'fitglm'* function allows an array of weights to be included as an additional input parameter. The vector of weights was constructed using the dispersion matrix first using the complement method and then the reciprocal method. The calculation of the two different weight vectors used to perform the FS is shown below:

- Complement method:

$$WEI = 1 - \frac{Err}{Err_{max}} \tag{2.29}$$

  Where the Err is the dispersion of the features that the algorithm is taking into account, the Err_{max} is the maximum dispersion value of the selected features.

- Reciprocal method:

$$WEI = \frac{|\frac{Delta}{Err}|}{max|\frac{Delta}{Err}|} \tag{2.30}$$

  Where Delta is the matrix of delta value of the considered features; the Err is the matrix of dispersion value of the same features, so it has the same size. The matrix: $|\frac{Delta}{Err}|$ is obtained by a point-by-point operation, and its maximum is obtained by vertically averaging the values to obtain an average dispersion for each feature.

The vector of weights must contain a number of elements equal to the number of observations so that the algorithm can give each observation a different relevance. For this, the WEI matrix was averaged horizontally, obtaining an average dispersion value between the features for each subject. From these two weighted FS, the features with the best performance were obtained in a similar way to the unweighted model.

2.5.4 Feature Validation

Validation of the features was done using the classification learner available in the APPs of Matlab R2020a. This interface allows loading a data matrix containing the features and their classes and manually selecting the features to be used for classification. In order to avoid overfitting errors, cross-validation was used, which divides the dataset into 5 folds and estimates the accuracy. The Classification Learner allows to choose the different classification model from the many available, including Logistic Regression. As input data, it is only possible to enter the feature matrix and not also the matrix of relative weights. Then the features obtained from the different FS performed will simply be selected a priori and used for classification. At the end of the model training, the relative Accuracy, confusion matrix, scatter plot and ROC curve of the model are available.

2.6 Relation between VHI and GIRBAS scales

Together with the recordings, general patient information such as age, disease status and other general characteristics were also provided. Ratings according to the GIRBAS and VHI scales for each record were also available. The GIRBAS scale is the rating scale used by speech therapists to assess the vocal quality of patients. The VHI, on the other hand, is a self-assessment made by the patients themselves. It was thus possible to compare the experts' perceptual assessment with the patients' self-assessment. The GIRBAS scale has a score from 0 to 4 where 0 indicates a normal condition and 4 a severe condition, the VHI on the other hand has a score from 0 to 120, where 120 indicates the worst situation. For both scales, a higher score indicates a worse situation. To make the comparison, delta values

derived this time from the subtraction $T_0 - T_1$ were used, so that positive delta values indicate an improvement. Since the GIRBAS scale consists of 6 different values for this analysis, the G and A values, which, on the advice of experts, are more relevant for patients with Multiple Sclerosis, are observed more carefully.

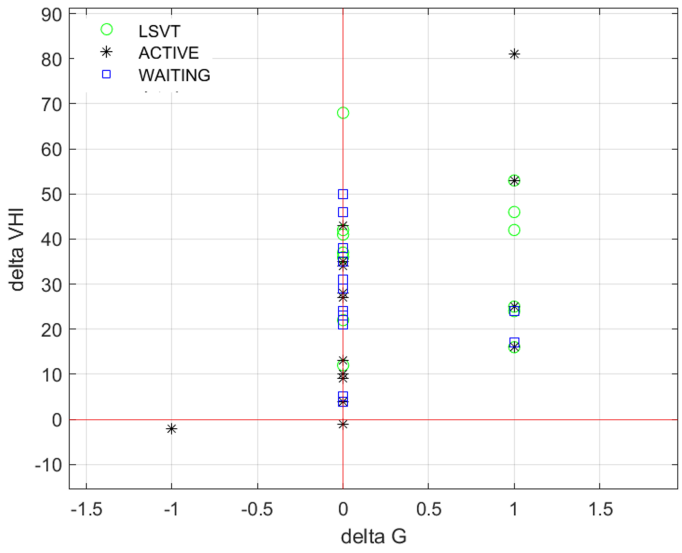

Figure 2.14: Scatter plot comparing self-reported VHI and perceptual G values (Girbas scale) of patients.

To visualise the results, the scatter plot (in Fig:2.14 and Fig:2.15) was used to compare two magnitudes on the two time-independent axes. If a value falls in the upper right quadrant, the perceptual evaluation and the self-assessment are consistent: the patient feels better after therapy (could be LSVT-LOUD, ACTIVE or WAITING) and the practitioner finds him/her improved. If, on the other hand, the value falls in the lower left-hand quadrant, then the two assessments are also consistent but in a negative sense, the patient is worse off and feels even worse. The other two quadrants, on the other hand, indicate inconsistency between the patient's feeling and the expert's assessment: in the bottom right-hand quadrant we find a patient who performs better but feels worse and in the top left-hand quadrant a patient who has got worse but feels better. In order to distinguish the type of rehabilitation therapy of the patients, different markers were used for the 3 therapy classes. However, it was observed that in almost all cases the VHI self-assessment score indicated an improvement both where the GIRBAS scale

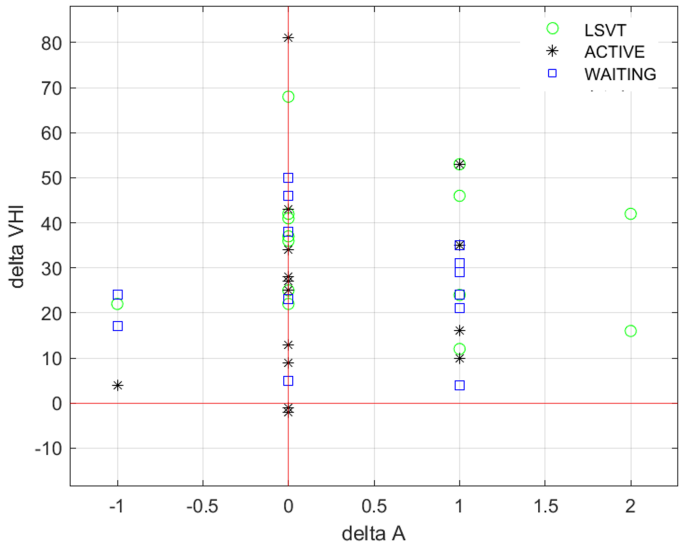

Figure 2.15: Scatter plot comparing self-reported VHI and perceptual A values (girbAs scale) of patients.

confirmed it and where it did not. To confirm the unreliability of the VHI scale, another available data set of laryngectomised patients was similarly investigated (using only the VHI scale in comparison to the GIRBAS scale). Here too, any positive or negative GIRBAS score was associated with positive VHI values, so this second observation reconfirmed the thesis, so that the VHI score will no longer be taken into consideration from here on.

2.7 Comparison between Extracted Features and GIRBAS scale

Another analysis conducted in this study concerns the relationship between the GIRBAS scale assigned by the experts and the result of parameter extraction. GIRBAS values, similarly, were converted into delta values, but according to the formula $T_1 - T_0$, in order to associate positive delta values with an improvement in vocal performance. The features used during the study are the delta values obtained from the subtraction of the extracted parameters at times T_1 and T_0 according to the formula $T_1 - T_0$. The extracted parameters are different and for some of them the increase corresponds to an improvement of the patients'

performance, while for others it corresponds to a worsening. Delta Features values were multiplied by a vector containing positive and negative unit values in order to allow direct association of delta Features values with GIRBAS delta values. Having made this change, the new positive delta features values represent an improvement, vice versa a deterioration. For the comparison, the scatter-plot was used, which makes it possible to relate two time-independent variables on the two axes of definition. For the comparison, we concentrated on observing the G (grade) and A (asthenic) parameters of the GIRBAS scale, which, according to expert advice, are the most significant for multiple sclerosis patients. With regard to the features, the features selected during the FS were observed (these were the features with the most class-related characteristics) and other features of particular clinical relevance. This assessment was made in a similar manner for both free speech and the vowel /a/. Observing the scatter plots in Fig: 2.17 and Fig: 2.16 is possible to observe the relationship between the A and G components of the GIRBAS scale and the features. As a visual reference one takes the axes of definition x and y and observes

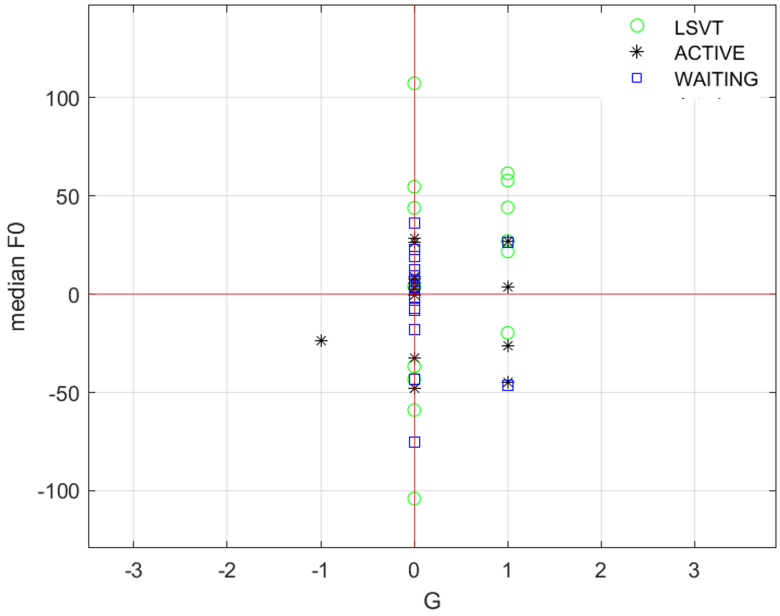

Figure 2.16: Relationship between features and component G of the GIRBAS assessment, in both cases positive values correspond to an improvement following therapy.

the consistency of the values: if a point is in the upper right or lower left quadrant

(resulting in x>0, y>0 and x<0, y<0) then the result of the extraction is consistent with the GIRBAS evaluation, positive and negative respectively. On the other hand, if the dot falls in the upper left quadrant (x<0, y>0), this means that according to the extraction algorithm the patient has improved while according to the experts he has worsened; conversely, if the dot is in the lower right quadrant (x<0, y>0), the patient has worsened according to the algorithm but improved according to the experts. From these scatter plots, it was possible to construct confusion matrices

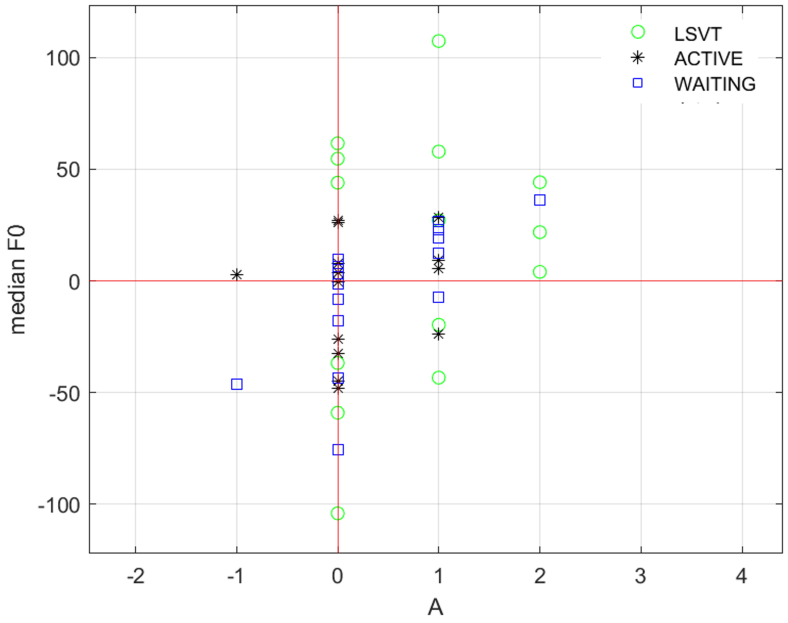

Figure 2.17: Relationship between features and component A of the GIRBAS assessment, in both cases positive values correspond to an improvement following therapy.

to compare the algorithm's evaluative capabilities with the experts' reliable ones. Two classes of patients were then constructed:

- Positive: patients who improved their vocal performance following therapy (Class: 2).
- Negative: patients who worsened their vocal performance following therapy (Class: 1).

For a correct interpretation it is important to remember that the health conditions

of the patients under investigation are different and Multiple Sclerosis is a neurodegenerative disease that can lead to more or less severe worsening over a short or longer period. Taking the neutral axes as a reference, four categories of patients were defined considering the experts' assessment as a real class:

- x>=0, y>=0 : True Positive
- x<0, y<0 : True Negative
- x>=0, y<0 : False Negative
- x<0, y>=0 : False Positive

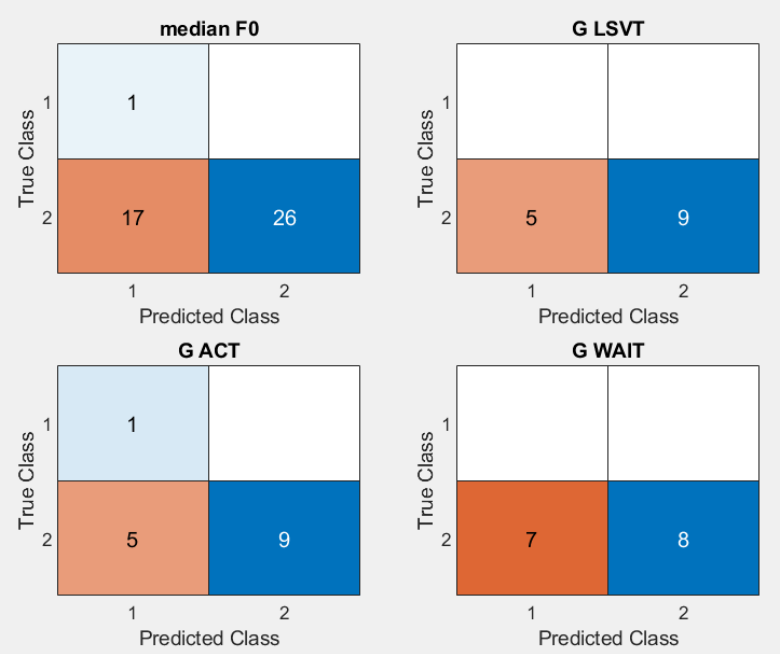

Figure 2.18: CM general (all patients) and partial (divided by therapy classes) where the true class is obtained from the GIRBAS (only G component) evaluation and the predicted is obtained from parameter extraction.

The general and partial Confusion Matrix (CMs) were thus obtained for each feature considered, Fig: 2.18 and 2.19; bottom right shows True Positives, bottom left False Negatives, top left True Negatives, top right False Positives. True results are assigned the colour blue, which intensifies with the number of elements, similarly, false results are associated with the colour orange. The general Confusion

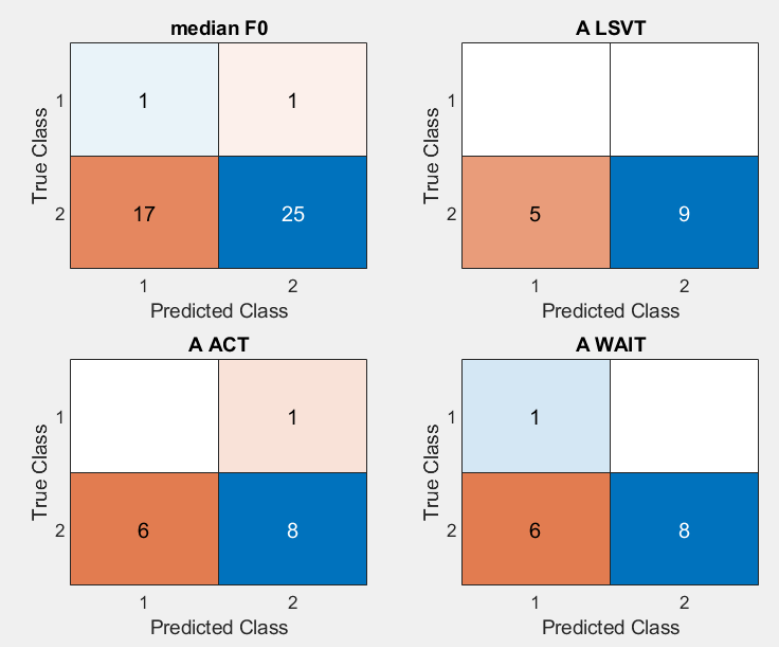

Figure 2.19: CM general (all patients) and partial (divided by therapy classes) where the true class is obtained from the GIRBAS (only A component) evaluation and the predicted is obtained from parameter extraction.

Matrix in the top left-hand corner is taken as the main reference, showing all patients regardless of therapy class. The other three CMs report partial class results to assess whether there are different inter-class trends. From these CMs, it was possible to calculate the evaluation metrics listed above to assess the performance of the extraction algorithm with reference to the GIRBAS reference evaluation. So by comparing the evaluation metrics of each feature, the best ones were chosen. In this case, however, there is no selection of features, but it is possible to identify the most robust and reliable features because they are the ones that most reflect and confirm the experts' assessment of the patients' vocal abilities.

Realistic Comparison

Observing the scatter plots, Fig: 2.17 and Fig: 2.16, it is possible to see that many points fall on the x-axis = 0, which corresponds to a delta G or A value of zero. That situation occurs when the experts assessed the patients' performance before and after therapy with a similar score. The possible values for the delta

characteristics, however approximate, are many more and it almost never happens that the value is equal to zero. This situation often occurs in view of the fact that the GIRBAS scale only assigns 5 different scores for each of the 6 parameters. Since the values are divided into the four types TP, TN, FP and FN using the axes as a boundary, the situations of greatest uncertainty and doubt are those in which the element straddles one of the two axes. The method used to split the elements divides the plane into 4 areas, the elements that fall on the x-axis=0 are considered positive (patients with improvement) and are included in the right areas of the plane using the expression >=0; this is justified by the fact that this is a degenerative disease and therefore a maintenance of the condition is still positive. Looking at the CMs and metrics (Fig:2.19 and Fig:2.18), it can be seen that the predominant error is due to FNs, i.e. patients who have improved but whom the algorithm considers to have worsened. As far as the method of assigning FNs is concerned, it cannot be ignored that this leads to innumerable errors, so it was decided to remove the elements with x=0 and y>0 from the dataset to create the realistic situation. Therefore, a realistic sub-dataset was created and a similar procedure was carried out. The CM matrix was observed and evaluation metrics calculated; finally, the results of the realistic case were compared with those of the original dataset. The scatter plots and general CMs (all patients belonging to the sub-dataset) of the new ideal situation for the G (Fig:2.20) and A (Fig:2.21) parameters of the GIRBAS scale, respectively, are shown in the figures.

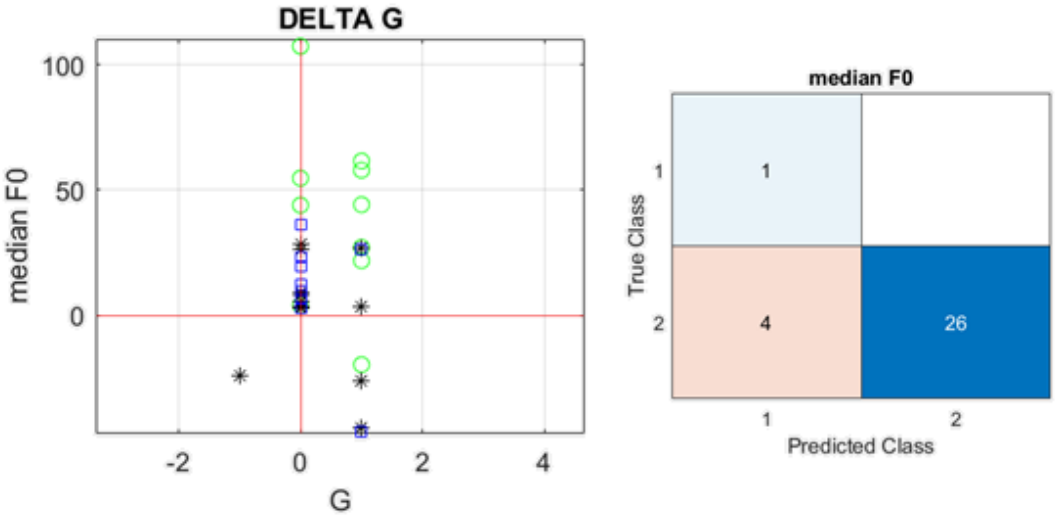

Figure 2.20: Ideal relation between delta figures and delta G.

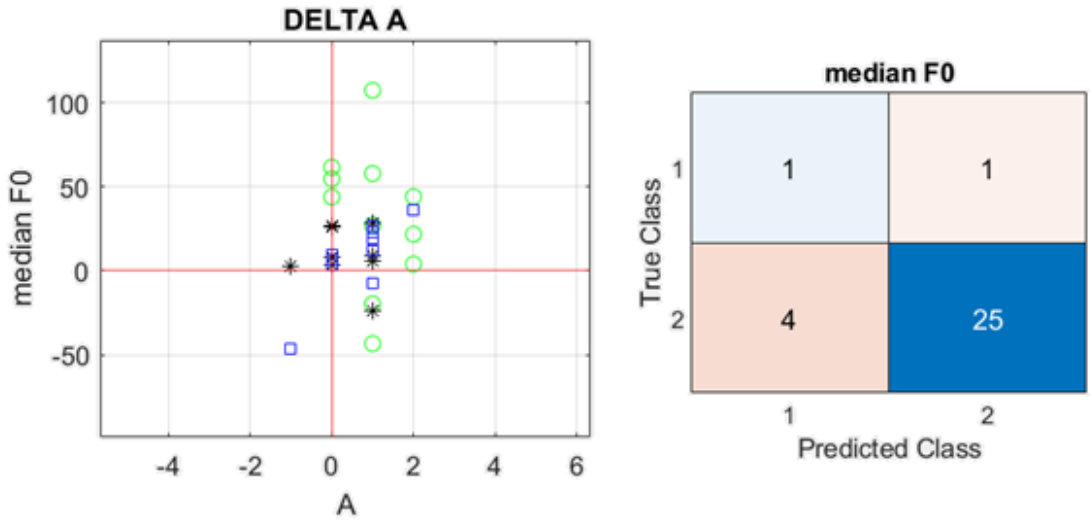

Figure 2.21: Ideal relation between delta figures and delta A.

Chapter 3

Results and Discussion

In this chapter, the results obtained from the different studies conducted on the extracted vocal parameters are discussed. In particular, the results of the comparisons between the vocal parameters obtained from recordings of multiple sclerosis (MS) patients and those obtained from recordings of healthy patients will first be illustrated. Next, the features obtained from the different feature selections conducted on the three pathological patient classes will be listed with their validated accuracies. Finally, the consistency of the experts' GIRBAS perceptual evaluation and that obtained from the voice parameters extracted from the algorithm will be discussed.

3.1 Pathological vs. Healty Results

In order to reliably compare the parameters relating to pathological patients and those relating to healthy patients, they were extracted using the same algorithm. The comparison was made between the absolute values of the parameters for healthy patients and those for all pathological patients before starting rehabilitation therapy (T_0) in order to consider the two most extreme and distinguishable situations and to identify which parameters differed the most in the two cases. Subsequently, it was decided to also include the comparison with pathological patients at time T_1 (after finishing therapy) divided into:

- Therapised patients, both with LSVT-LOUD and with ACTIVE therapy, to

see how close they are to the healthy ones after therapy.

- Untherapised patients, to observe the differences with the therapised patients.

With this further comparison, it was also possible to see which features are most affected by the therapy and thus which features improved most and came closest to the value of the healthy ones. In order to choose the best features, the graphs of the mean values of the 4 categories of patients under investigation (Healthy, Pathological T_0, Therapised T_1, Non-therapised at T_1) were observed, on the graphs also the dispersions relative to the 4 broad $\pm 2\sigma$ groups (thus including 95.45% of the cases) are shown. In the case where between healthy and pathological patients at T_0 the dispersion bands do not overlap, the feature is selected. At this point, the relative values from pathological patients at time T_1 treated and untreated are also observed.

3.1.1 Vowel /a/ Results

For the vowel /a/ 17 parameters were identified as significantly different from a statistic point of view. The parameters shown in Tab:3.1, where the mean value and standard deviation relative to the 4 classes of patients under observation are shown, on the first two columns are the feature names and their unit of measurement. Looking at the values of the pathological patients following therapy, one can observe an improvement in condition compared to the pre-therapeutic condition, while remaining closer and with overlapping dispersion bands to the values of the pathological pre-therapeutic patients; an approximation to the value for healthy patients can be observed in all the parameters listed in Tab:3.1 except for : vF0, std and 95° prc of f_0; for which the post-therapy situation seems to have worsened. Another thing that can be seen from the graphs is that often the value of non-therapy patients is similar to or even higher than the values of the healthy patients compared to the treated patients; this apparently abnormal result could be due to the variable pathological condition of the individual patients.

Feature	m.u.	SM T0 (44) Tutti	SM T1 (29) LSVT+ACTIVE	SM T1 (15) WAITING	Sani (56)	
CPPS 5° prc	dB	12,85	13,70	13,46	14,52	mean
		0,28	0,35	0,56	0,25	std
CPPS std	dB	1,59	1,50	1,55	1,32	mean
		0,05	0,04	0,05	0,03	std
f0 range	Hz	42,43	40,16	36,77	14,52	mean
		4,28	4,22	7,50	1,37	std
f0 std	Hz	5,71	6,20	4,80	1,85	mean
		0,73	1,10	1,00	0,19	std
f0 95° prc	Hz	198,30	211,30	194,60	167,00	mean
		7,15	10,08	13,21	7,07	std
HNR 5° prc	dB	8,27	9,68	9,91	11,88	mean
		0,44	0,54	0,63	0,42	std
HNR mode	dB	12,11	13,86	13,93	15,43	mean
		0,54	0,54	0,75	0,47	std
HNR median	dB	14,49	15,89	16,16	17,33	mean
		0,42	0,47	0,54	0,47	std
HNR mean	dB	15,31	16,70	16,93	18,19	mean
		0,41	0,48	0,55	0,49	std
Vam	%	23,70	22,70	25,20	15,80	mean
		0,90	1,30	2,30	1,00	std
apq	%	4,84	4,18	4,02	2,15	mean
		0,35	0,45	0,49	0,18	std
shimmer	dB	0,51	0,42	0,39	0,24	mean
		0,04	0,05	0,04	0,02	std
local shimmer	%	5,82	4,73	4,38	2,71	mean
		0,43	0,50	0,42	0,22	std
vfo	%	2,93	3,04	2,40	1,11	mean
		0,33	0,51	0,33	0,11	std
ppq	%	0,38	0,30	0,25	0,22	mean
		0,04	0,06	0,03	0,01	std
rap	%	0,38	0,26	0,22	0,20	mean
		0,05	0,04	0,03	0,02	std
local jitter	%	0,63	0,45	0,37	0,35	mean
		0,08	0,08	0,04	0,03	std

Figure 3.1: The table shows the 17 parameters that are statistically independent (95.45%) when comparing MS patients (T_0) vs. Healthy

3.1.2 Free speech Results

In the case of free speech 10 parameters are result significantly different from a statistic point of view (dispersion bar set at 2σ). Tab:3.2 shows the mean values and standard deviations of the 4 classes of patients under investigation, in the first two columns are the name and relative unit of the parameter. The selected parameters are the std and range of the HNR, while for the CPPS all descriptive statistics except kurtosis are selected. When looking at the values for the CPPS statistics, it can be seen that the mean, median and mode are lower in pathological patients than in healthy ones, as might be expected; moreover, the values increase in a slightly higher value following therapy (improvement in the patient's condition). On the other hand, the range and std of the CPPS are higher in healthy patients than in pathological patients, which is not consistent with what is expected; 5th and 95th prc of the CPPS confirm the lower dispersion of the values in pathologists than in healthy patients. In general, it is observed that the vowel /a/ presents many more statistically independent and consequently significant parameters for the distinction between pathological and healthy patients, of which many more parameters are positively enhanced by voice therapy.

3.2 Logistic Regression Result

Logistic regression was used to find the most significant features for the classification of the 3 classes of pathological patients. The delta values of the extracted parameters were used as features by going to subtract the values at T_0 from those at T_1. Logistic Regression is a type of binary classification so three features selection was made for both vowel /a/ and free speech. After identifying the best feature, pair, triplet or quatern of features, they were validated using the Classification Learner on Matlab APP using the 5 fold cross validation option. In the validation phase, we went to observe the accuracy values of the logistic regression but also tried the other classifiers available on the APP and reported the accuracies that were greater than that of the logistic regression and the corresponding classifier with which it was obtained. In the case of the vowel /a/, similar tests were also carried out in the two cases of weighted logistic regression, and the results obtained were similarly validated.

Feature	m.u.	SM T0 (39) Tutti	SM T1 (24) LSVT+ACTIVE	SM T1 (14) WAITING	Sani (47)	
HNR	dB	5,40	5,7	5,6	6,00	mean
std		0,14	0,18	0,3	0,05	std
HNR	dB	33,60	33,2	34	38,00	mean
range		1,22	1,36	2,6	0,61	std
CPPS	dB	11,05	11,23	11,3	12,4	mean
mean		0,22	0,27	0,5	0,08	std
CPPS	dB	11,240	11,480	11,62	13,140	mean
median		0,28	0,35	0,65	0,11	std
CPPS	dB	11,07	11,74	12,42	14,79	mean
mode		0,63	0,75	0,98	0,3	std
CPPS	dB	3,510	3,550	3,6	4,160	mean
std		0,072	0,09	0,11	0,03	std
CPPS	dB	18,03	18,22	18,41	19,93	mean
range		0,23	0,36	0,43	0,1	std
CPPS	dB	5,450	5,460	5,32	5,300	mean
5° prc		0,06	0,09	0,17	0,03	std
CPPS	dB	16,43	16,63	16,7	18,2	mean
95° prc		0,24	0,32	0,49	0,09	std
CPPS	dB	-0,050	-0,100	-0,14	-0,330	mean
skewness		0,06	0,06	0,12	0,02	std

Figure 3.2: The table shows the 4 parameters that are statistically independent (95.45%) when comparing MS patients (T_0) vs. Healthy

3.2.1 Vowel /a/ Results

As mentioned before 3 features selection and validation were performed for each combination of the 3 classes, the results are shown in the table, on the left the results in terms of accuracy not validated and on the right the results with validation; any results with higher accuracy from other classification methods are shown in parentheses.

LSVT vs. ACTIVE

Tables in green show the binary classification results between patients treated with LSVT-LOUD (LSVT: 14 patients) and patients treated standard therapy (ACT: 15 patients), in the first table the results related to unweighted classification, Tab: 3.3, in the second table the results related to the weighted classification with the complement method, Tab:3.4, in the third table the results related to the weighted classification with the reciprocal method, Tab:3.5, lastly the table with the results of classification without performing features selection Tab:3.6. Referring to the

LSVT (14 pz) vs ACT (15 pz)	Classification Accuracy /a/	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID
skewness CPPS	72,4%	72,4%
vF0	75,9%	72,4%
95° prc HNR		
apq	**86,2%**	**72%**
mode HNR		
skewness CPPS		
PARAMETERS	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.3**	

Figure 3.3: Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

table 3.3 it can be seen that the highest accuracy, in the unvalidated case, occurs for the three features (86.2%); in each case all validated accuracies exceed 70% and coincide almost in the same value of 72%. The features that are selected are two stability parameters (vF0 - apq), two statistics of the HNR (mode and 95° prc) and the skewness of the CPPS. The p value of the model is set to the value 0.3, the minimum value for which at least one feature triplet is obtained. Referring to the table 3.4 it can be seen that the highest accuracy, in the unvalidated case, occurs for the three features (79.3%); the validated accuracies are on the order of 70%, but it can be seen that using KNN instead of LR in the validation phase would

LOGISTIC REGRESSION (Complement-weighted)		
Features	NO VALID	5-fold cross VALID
skewness CPPS	72,4%	72,4%
vF0	75,9%	72,4%
95° prc HNR		
ppq	**79,3%**	**69% (KNN:79,3%)**
mode HNR		
skewness CPPS		
PARAMETERS	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.4**	

Figure 3.4: Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

result in 79.3% by increasing the accuracy by 10 percentage points. The selected features turn out to be very similar to the unweighted classification case differing only in the ppq parameter instead of apq. The p value of the model in this case was increased to the value 0.4, indicative of more correlated features. Looking

LOGISTIC REGRESSION (Reciprocal-weighted)		
Features	NO VALID	5-fold cross VALID
95° prc fo	48,3%	65,5% (SVM:69%)
std RMS	62,1%	62,1%
kurtosis RMS		
range RMS	**75,9%**	**72,4% (SVM, ensamble:75,9%)**
95° prc HNR		
std f0		
PARAMETERS	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.5**	

Figure 3.5: Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

at the table 3.5 we see generally lower values of accuracy in both the validated and non-validated cases, although in the case of the three features we arrive at accuracy values comparable to the previous two cases; again we can see that using other classification methods (SVM or ensamble) to perform validation increases the accuracy value up to the value of the non-validated case (75.9%). In this case the p value of the model was raised up to 0.5; this confirms the worst classification results. The validated classification results without features selection are much

NO feature selection	Classification Accuracy /a/
	5-fold cross VALID
all features	58,6% (KNN: 72,4%)

Figure 3.6: validated classification results without features selection

lower, using LR only 58% accuracy is obtained while using KNN reaches 72.4%, similar to the validated cases with features selection. Therefore, it can be inferred that the KNN might be a good classifier for this type of data. In general, it is also noted that weighted grading does not lead to particular improvements, but that it can also sometimes be pejorative.

LSVT vs. WAITING

The case in which the best results in terms of accuracy are obtained with model validation is the classification of LSVT and WAITING subjects, in the unweighted case the best results are obtained in the case of the three features, in which values in the region of 80% are obtained. Looking at the Tab: 3.7 it can be seen that the selected features are different statistics of the HNR (95° prc, 5° prc and range), the stability parameter vF0 and the range of the RMS, which is, however, taken into little consideration due to the inconsistent acquisition modes. The p-value in this case is set at 0.2, one of the lowest values among the tests performed, confirming the best results. It is also noted that SVM might be a better model than logistic regression in the validation phase. Observing Tab: 3.4 we can see that by introducing feature weighting the validated results are better, with accuracies always above 70% up to almost 80% in the case of three features, where again

LSVT (14 pz) vs WAIT (15 pz)	Classification Accuracy /a/	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID.
95° prc HNR	65,5%	58,6% (NB,SVM:69%)
vF0	79,3%	69% (QD,SVM:72,4%)
5° prc HNR		
range HNR	**82,7%**	**79,3% (SVM:82,8%)**
95° pcr F0		
range RMS		
PARAMETRI	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.2**	

Figure 3.7: Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

LOGISTIC REGRESSION (Complement-weighted)		
Features	NO VALID	5-fold cross VALID
skewness CPPS	72,4%	72,4%
vF0	79,3%	75,9% (LD:79,3%)
95° prc HNR		
range HNR	**79,3%**	**79,3% (SVM:82,8%)**
95° prc fo		
range RMS		
PARAMETRI	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.4**	

Figure 3.8: Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

the SVM turned out to be an improving method. In this case the p-value was increased to 0.4 indicating less distinguishable features. The best combination,

that of the three features, is consistent with the unweighted case and is always the best. Looking at Tab:3.9 we can see that the features selected are the same

LOGISTIC REGRESSION (Reciprocal-weighted)		
Features	NO VALID	5-fold cross VALID
95° prc HNR	62,1%	58,6% (NB,SVM:69%)
vF0	75,9%	75,9% (LD:79,3%)
HNR (95° prc)		
range HNR	**79,3%**	**79,3% (SVM:82,8%)**
95° prc fo		
range RMS		
PARAMETRI	Min corr= 0 // Max corr= 0.7 Min pv= 0.05 // **Min pv mod= 0.6**	

Figure 3.9: Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

in all cases considered, so in this case we can say that the weighting has had no influence on feature selection and the results are similar. What does change is the value at which the p-value is set, which is much higher, so we can deduce that this weighting is not improving. Not doing feature selection (Tab:3.10) leads to very

NO feature selection	**Classification Accuracy** *Val*
	5-fold cross VALID
all features	**48,3% (SVM,KNN: 65,5%)**

Figure 3.10: validated classification results without features selection

low results, less than 50 per cent, and LR is certainly not the preferred method, in which case SVM is certainly advantageous, confirming the cases seen above.

ACT (15 pz) vs WAIT (15 pz)	Classification Accuracy /a/	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID.
kurtosis RMS	63,3%	60,0%
vF0 %	**70,0%**	**73,3%**
5° prc RMS		
Jitter	63,3%	53,3% (SVM:66,7%)
median HNR		
5° prc CPPS		
apq	66,7%	46,7% (SVM:60%)
Vam		
std fo		
std RMS		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.7**	

Figure 3.11: Unweighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

ACTIVE vs. WAITING

In general, the least distinguishable case turns out to be the one between ACTIVE and WAITING, observing the Tab:3.11 we immediately notice how the results are much lower than the previous cases, the best case turns out to be the one that uses two features and arrives at around 70%. In general, many stability parameters are selected (VF0, apq, Vam, Jitter), the std of the f_0 the median of the HNR, the 5th prc of the CPPS and the statistics of the RMS are not taken into much consideration. In this case, the algorithm also manages to select 4 features but sets the p-value to 0.7, which is a dicisely high value. It should be noted that in this case, too, the SVM proves itself to be a vantage model for this type of classification.

Looking at the Tab:3.12 we can see that the weighting of the features leads to decidedly worse results, arriving in the best case at an accuracy of 60 % in

LOGISTIC REGRESSION (Complement-weighted)		
Features	NO VALIDATION	5-fold cross VALID.
kurtosis RMS	60,0%	**60,0%**
kurtosis HNR	**63,3%**	50%(SVM:53,3%)
kurtosis RMS		
rap	60,0%	50%(KNN:53,3%)
mean HNR		
5° prc CPPS		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.5**	

Figure 3.12: Complement-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

the validated case, i.e. 10 percentage points less than the unweighted case. In this case the p-value is a little lower, 0.5, and in fact one does not get to select 4 features as in the previous case. The highest accuracy is obtained using only one feature (first situation) which is the kurtosis of the RMS, which we had considered unreliable due to acquisition problems; certainly this situation is worse than the unweighted LR. The latter case in Tab: 3.13 turns out to be extremely worse, this time the best accuracy is 53.3%, with other classification methods it could be as high as 60% but still remains considerably low, and in this case the p value is again set to 0.7. Certainly feature weighting with the reciprocal method is not an advisable method for this type of data. Again, the results of classification without feature selection are low, Tab 3.14, about 45%, but consistent with the previous case of reciprocal-weighted classification. In general, in this third case of binary classification (ACTIVE vs. WAITING), it can be said that feature weighting is pejorative.

LOGISTIC REGRESSION (Reciprocal-weighted)		
Features	NO VALIDATION	5-fold cross VALID.
kurtosis RMS	56,7%	60,0%
range HNR	63,3%	**53,3%(Tree,NB:60%)**
kurtosis RMS		
Vam	**70,0%**	50%(NB:66,7%)
mean RMS		
std RMS		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.7**	

Figure 3.13: Reciprocal-weighted classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

NO feature selection	**Classification Accuracy *tal***
	5-fold cross VALID
all features	**46,7%% (ensamble:56,7%)**

Figure 3.14: validated classification results without features selection

3.2.2 Free speech Results

For the classification of patients in the case of free speech, it was not possible to carry out feature selection weighing but only simple feature selection as only one repetition was available. We therefore proceeded in a similar way to the case of the vowel /a/ by first carrying out the features selection using the combinatorial algorithm based on LR and then went on to validate the model on the Classification Learner (Matlab APP's) going on to identify, in addition to LR, other more advantageous classification methods.

LSVT vs. ACTIVE

Looking at the Tab:3.15, it can be seen that in the first two cases the validated accuracy always exceeds 75 %. The best case turns out to be the second, in which are used the 5° prc of CPPs and the V/uV parameter. Using alternative methods, such as the Decision Tree Classifiers, values above 92 % are obtained. The selected features are the range and the median of the HNR, 5° prc and range of CPPS and the V/uV parameter. In this case, the p-value was set at 0.2, which is a lower value compared to the case of the vowel /a/. The classification results without feature

LSVT (8 pz) vs ACT (16 pz)	Classification Accuracy ELOQUIO NEW	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID
5: HNR (range)	75,0%	75% (SVM, tree, NB: 83,3%)
33: CPPS (5° prc)	**83,3%**	**79,2%**
37: VluV		
2: HNR (median)	79,2%	58,3% **(Tree: 91,7%)**
32: CPPS (range)		
33: CPPS (5° prc)		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.2**	

Figure 3.15: Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

selection, Tab: 3.16, turn out to be extremely poor, with the LR achieving only 45.8%, whereas using other methods such as Ensemble or KNN yields higher results, 70.8%. Again, LR does not seem to be a preferable method in the validation phase.

LSVT vs. WAITING

Also in the case of speech, the classification between LSVT and WAITING patients turns out to be the most accurate, Tab: 3.17. In this case, significantly higher

NO feature selection	Classification Accuracy ELOQUIO NEW
	5-fold cross VALID
all features	45,8% **(KNN, Ensemble 70,8%)**

Figure 3.16: validated classification results without features selection

results are obtained, reaching values of the order of 80 % in the case validated with LR; while using other methods, such as SVM, it is up to 91 %. The selected features are the mode, the mean and the 95° prc of CPPS; ths std and the 95° prc of f_0 and the mean of HNR. The p value is set to 0.1, the lowest value in all the features selection performed in this study. The results in Tab:3.18 are obtained not

LSVT (8 pz) vs WAIT (15 pz)	Classification Accuracy ELOQUIO NEW	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID
30: CPPS (mode)	81,8%	81,8% (KNN, SVM 86,4%)
16: f0 (95° prc)	**90,9%**	**77,3% (SVM 90,9%)**
28: CPPS (mean)		
1: HNR (mean)	**90,9%**	**81,8%**
13: f0 (std)		
34: CPPS (95° prc)		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.1**	

Figure 3.17: Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

carrying out features selection. Also in this case accuracy are worse. Using the Naive Bayes Classifier the accuracy is higher and reaches the 72.3 per cent.

ACTIVE vs. WAITING

As in the case of the vowel /a/ the classification with the poorest results is again the one between ACTIVE and WAITING patients. Referring to the Tab: 3.19, the

NO feature selection	Classification Accuracy ELOQUIO NEW
	5-fold cross VALID
all features	**50% (NB 72,3%)**

Figure 3.18: Validated classification results without features selection

cases validated with LR, accuracy of 73.3 % are reached in all the cases except for the first selection, in which is used only one features.In this case the LR results to be the best classifier even in validation phase. The selected features are the mode, the mean of the HNR, the mode and the std of CPPS, the mode and the 5° prc of RMS. The p-value in this case is set to 0.5 that is considered an high value; on the other hand it allows to perform the selection of 4 features. The results in

ACT (16 pz) vs WAIT (15 pz)	Classification Accuracy ELOQUIO NEW	
	LOGISTIC REGRESSION	
Features	NO VALIDATION	5-fold cross VALID
3: HNR (mode)	60,0%	60% (SVM 63,3%)
1: HNR (mean)	80,0%	73,3%
30: CPPS (mode)		
1: HNR (mean)	80,0%	73,3%
21: RMS (mode)		
24: RMS (5° prc)		
1: HNR (mean)	**80,4%**	**73,3%**
3: HNR (mode)		
30: CPPS (mode)		
31: CPPS (std)		
PARAMETERS	Min corr= 0 # Max corr= 0.7 Min pv= 0.05 # **Min pv mod= 0.5**	

Figure 3.19: Classification results: on the right the name of the features used, in the middle the accuracies without validation obtained, on the left the accuracies obtained from validation, at the bottom the parameters used by the combinatorial algorithm for feature selection

Tab: 3.20 are obtained not performing feature selection. Using the LR classifier accuracy of 50% is reached; this confirms the outcomes previously obtained by not implementing FS, Fig: 3.16 and Fig: 3.18. Using Ensemble Classifier for the validation higher accuracy are obtained, 67.7%.

NO feature selection	Classification Accuracy ELOQUIO NEW
	5-fold cross VALID
all features	50% **(Ensemble 67,7%)**

Figure 3.20: validated classification results without features selection

As expected, the classification that achieves the best results is that between LSVT and WAITING patients. In general, Logistic Regression is a good method for performing Features Selection. However, to validate the model, it was observed that other classification methods might be more advantageous.

3.3 VHI vs. GIRBAS scale Results

For each recording, both the experts' perceptual assessment (GIRBAS scale) and the patients' self-assessment (VHI scale) were available. The delta values of the two assessments (positive values indicate improvement and negative values a worsening) were compared on a dot plot. What was observed is that patients almost always tend to increase their scores following therapy and thus consider themselves to have improved in vocal performance. This makes the VHI assessment inconsistent but certainly indicates that the therapy has given hope and increased psychological confidence to the patients, which is very important when dealing with degenerative diseases.

3.4 Extracted Features vs. GIRBAS scale Results

Girbas delta values were also compared with feature delta values. To facilitate the evaluation, some delta feature values were changed in sign to obtain positive

values in the case of improvement and negative values in the opposite case. For each feature, a scatter plot was obtained in which the x and y axes divide the space into 4 regions from which the TP, TN, FP and FN were calculated. The confusion matrices with the corresponding evaluation metrics were then obtained. In particular, accuracy, precision, sensitivity and specificity were looked at first on all patients and then by class (LSVT, ACTIVE, WAITING).

3.4.1 Vowel /a/ Results

The average values for accuracy, precision, sensitivity and specificity are shown below, on the left column the values in relation to parameter G (Girbas scale), on the right the values in relation to parameter A (Girbas scale). The first row shows the total values, i.e. for all patients, the other rows the partial values for the three therapy classes under consideration. Looking specifically at Tab: 3.1 and 3.2, generally low values are observed, around 55%, except for LSVT patients, which reach almost 63%.

G (Girbas scale)	
Patient	**Accuracy**
TOTAL	56%
LSVT	63%
ACTIVE	56%
WAITING	50%

Table 3.1: Average accuracy of the relationship between the G parameter (GIRBAS scales) and extracted parameters.

A (Girbas scale)	
Patient	**Accuracy**
TOTAL	56%
LSVT	63%
ACTIVE	51%
WAITING	55%

Table 3.2: Average accuracy of the relationship between the A GIRBAS parameter (scales) and extracted parameters.

Looking instead at Tabs: 3.3 and 3.4 one can immediately notice for precision much higher values, between 95 % and 100 %; while for sensitivity (3.5 - 3.6) and specificity (3.7 - 3.8) the values return low. Thus, there is a glaring imbalance between precision and sensitivity. Reconstructing the formulae defining precision, sensitivity and specificity and unitaly observing the confusion matrix from which they were then calculated, one can see that a large part of the errors can be attributed to the presence of many false negatives (FN).

G (Girbas scale)	
Patient	**Precision**
TOTAL	98%
LSVT	100%
ACTIVE	96%
WAITING	100%

Table 3.3: Average precision of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Precision**
TOTAL	96%
LSVT	100%
ACTIVE	92%
WAITING	98%

Table 3.4: Average precision of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

G (Girbas scale)	
Patient	**Sensitivity**
TOTAL	56%
LSVT	63%
ACTIVE	55%
WAITING	50%

Table 3.5: Average sensitivity of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Sensitivity**
TOTAL	56%
LSVT	63%
ACTIVE	53%
WAITING	53%

Table 3.6: Average sensitivity of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

Looking at the scatter plots from which the confusion matrices were derived, it can be seen that many points (representative of individual patients) lie on the x=0 axis, i.e. delta G is zero. If these values are associated with y>0 then they are considered true positives (TP), if the associated values are y<0 then they are considered false negatives (FN); it therefore happens that patients who have been assessed in a similar way, before and after therapy, by the specialist but who, according to the algorithm, have had a worsening of their condition fall among the FN. This kind of error is partly due to the poor resolution of the GIRBAS scale, which is defined on 5 different values, where 0 represents a healthy item and 4 a very serious situation; small or medium variations are indefinable and this was also confirmed by the experts at Don Gnocchi.

G (Girbas scale)	
Patient	**Specificity**
TOTAL	60%
LSVT	0%
ACTIVE	60%
WAITING	0%

Table 3.7: Average specificity of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Specificity**
TOTAL	56%
LSVT	0%
ACTIVE	28%
WAITING	85%

Table 3.8: Average specificity of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

Realistic Result

In order to get an idea of what this evaluation might look like by eliminating the problem just described, it was decided to repeat the evaluation according to the metrics by eliminating from the calculation those patients who had x=0 and y<0 values. This time, however, only the overall values were calculated and not the therapy class partials.

G (Girbas scale)	
ACCURACY	80%
PRECISION	98%
SENSITIVITY	81%
SPECIFICITY	60%

Table 3.9: Evaluation metrics averaged over all features, in relation to parameter G, for all patients under examination.

A (Girbas scale)	
ACCURACY	77%
PRECISION	96%
SENSITIVITY	78%
SPECIFICITY	56%

Table 3.10: Evaluation metrics averaged over all features, in relation to parameter A, for all patients under examination.

By making this change, the average accuracy rises by about 20 percentage points to around 80%, the precision remains high in any case, the sensitivity increases considerably to around 80%, compared to 58% in the previous case; the specificity, although low, remains constant, not being affected by this change (Tab: 3.9 and 3.10).

3.4.2 Free speech Results

In the case of free speech, we behaved in exactly the same way; it can be observed that the results obtained have more or less the same order of magnitude. In fact the metrics are affected by the same type of error due to the excess of False Negative.

G (Girbas scale)	
Patient	**Accuracy**
TOTAL	51%
LSVT	59%
ACTIVE	52%
WAITING	45%

Table 3.11: Average accuracy of the relationship between the G parameter (GIRBAS scales) and extracted parameters.

A (Girbas scale)	
Patient	**Accuracy**
TOTAL	50%
LSVT	59%
ACTIVE	48%
WAITING	48%

Table 3.12: Average accuracy of the relationship between the A parameter (GIRBAS scales) and extracted parameters.

G (Girbas scale)	
Patient	**Precision**
TOTAL	95%
LSVT	100%
ACTIVE	89%
WAITING	100%

Table 3.13: Average precision of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Precision**
TOTAL	100%
LSVT	69%
ACTIVE	90%
WAITING	98%

Table 3.14: Average precision of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

Realistic Result

In Tab:3.19 and Tab 3.20 are reported the evaluation metrics after the removal of the error due to the element with x=0 and y<0. The metrics report similar result as the case of the vowel /a/: the average accuracy increase, the precision remains high in any case, the sensitivity increases, compared to 58% in the previous case; the specificity, remains constant, not being affected by this change.

G (Girbas scale)	
Patient	**Sensitivity**
TOTAL	50%
LSVT	59%
ACTIVE	51%
WAITING	45%

Table 3.15: Average sensitivity of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Sensitivity**
TOTAL	50%
LSVT	59%
ACTIVE	49%
WAITING	46%

Table 3.16: Average sensitivity of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

G (Girbas scale)	
Patient	**Specificity**
TOTAL	59%
LSVT	0%
ACTIVE	59%
WAITING	0%

Table 3.17: Average specificity of the relationship between the G parameter (GIRBAS scale) and extracted parameters.

A (Girbas scale)	
Patient	**Specificity**
TOTAL	53%
LSVT	0%
ACTIVE	38%
WAITING	67%

Table 3.18: Average specificity of the relationship between the A parameter (GIRBAS scale) and extracted parameters.

G (Girbas scale)	
ACCURACY	75%
PRECISION	95%
SENSITIVITY	78%
SPECIFICITY	51%

Table 3.19: Evaluation metrics averaged over all features, in relation to parameter G, for all patients under examination.

A (Girbas scale)	
ACCURACY	71%
PRECISION	95%
SENSITIVITY	72%
SPECIFICITY	47%

Table 3.20: Evaluation metrics averaged over all features, in relation to parameter A, for all patients under examination.

Chapter 4

Conclusions

In this thesis work, three different studies were conducted on vocal parameters extracted from pathological patients and healthy subjects.
The first study relates parameters from pathological subjects, before voice therapy (T0), to values from healthy subjects. All parameters are extracted using the same algorithm to make the values comparable. Looking also at the value of the pathological subjects after voice therapy (both LSVT-LOUD and ACTIVE at T1), it is possible to observe how much the voice therapy positively influenced that parameter by bringing it closer to the characteristic value of healthy subjects. The parameters that best discriminate between pathological and healthy patients are: CPPS (5° pcr, std), f_0 (std, range, 95° prc), HNR (mean, median, mode, 5° prc), Vam, APQ, Shimmer db, Shimmer %, Vfo, PPQ, RAP, local Jitter, for the vowel /a/, Tab:3.1; HNR (std, range), CPPS (mean, median, mode, std, range, 5° prc, 95° prc, skewness) for free speech, Tab:3.2. The selected parameters confirm what could be expected. No RMS statistics are selected, as the data acquisition of pathological patients was carried out under amplitude uncontrolled conditions. Increasing vocal amplitude is the primary goal of voice therapy, therefore a future study conducted on consistent amplitude data could be very interesting. Men and women have different vocal frequency ranges, with higher average values in women. The parameter f0 is selected using dispersion statistics (range and std of f_0) and never using central tendency statistics. A subsequent study could be conducted by dividing male patients from females in the observation of f_0 statistics.

The second study conducted concerned only pathological subjects divided into the three therapy classes. The delta values of the extracted parameters between times T_0 and T_1 were considered as subject features. Three binary classifications were conducted for the three therapy classes of patients, with the aim of finding the features for which the classes were most distinguishable. In general, no very high classification results were obtained, around 80%, but this is not surprising as the data were very similar: patients with the same disease are not differentiated by severity of the disease. The best result of classification are obtained for the free speech, specifically in the case of LSVT vs. WAITING (non-therapised patients). The best validated accuracy, around 82%, occurs in two cases: the single feature CPPS mode and for the triplet of features HNR mean, fo std, CPPS 95° prc. Also other methods of classification were used in validation phase, it was often noted that other classification methods such as Dispersion Trees, SVM and KNN can increase the performance. In fact the best classification accuracy is reached in free speech classification between LSVT and ACTIVE patients. In this case an accuracy of 92% is reached using 3 features (HNR median, CPPSrange, CPPS prc) and the Dispersion Tree for validation. In general, the Dataset used is too small to achieve good classification results. If a larger Dataset was available, it would be possible to differentiate patients also by severity of vocal condition (using the Girbas evaluation).

The last analysis conducted verifies that the algorithm used for feature extraction returns values that are consistent with what the experts assess in terms of improvement or worsening of vocal performance. The problem encountered is in the poor resolution of the rating scale, which makes it impossible to assess small changes. In this sense, the algorithm could be used as a support tool for experts during perceptual evaluation. Moreover the algorithm could be used also to make an early diagnosis by overcoming the human ear in term of resolution.

Bibliography

[1] Gustavo Xavier Andrade Miranda. «Analyzing of the vocal fold dynamics using laryngeal videos». PhD thesis. Telecomunicacion, 2017 (cit. on p. 2).

[2] Tech. rep. available online. URL: https://www.guwsmedical.info/human-body/components-and-subdivisions-of-the-human-respiratory-system.html (cit. on p. 3).

[3] Tech. rep. available online. URL: https://dysphonia.org/your-journey/how-voice-works/ (cit. on p. 4).

[4] 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).

[5] Valeria Crispiatico, Cinzia Baldanzi, Arianna Napoletano, Laura Tomasoni, Francesca Tedeschi, Elisabetta Groppo, Marco Rovaris, Chiara Vitali, and Davide Cattaneo. «Effects of voice rehabilitation in people with MS: A double-blinded long-term randomized controlled trial». In: *Multiple Sclerosis Journal* 28.7 (2022), pp. 1081–1090 (cit. on p. 6).

[6] Anna Miles, Marie Jardine, Felicity Johnston, Martin de Lisle, Philippa Friary, and Jacqui Allen. «Effect of Lee Silverman Voice Treatment (LSVT LOUD®) on swallowing and cough in Parkinson's disease: A pilot study». In: *Journal of the Neurological Sciences* 383 (2017), pp. 180–187. ISSN: 0022-510X. DOI: https://doi.org/10.1016/j.jns.2017.11.015. URL: https://www.sciencedirect.com/science/article/pii/S0022510X17344490 (cit. on p. 6).

[7] Studio medico Tarzaniello. *Valutazione Percettiva della Voce*. Tech. rep. available online. IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition, 2012. URL: http://www.tanzariello.it/index.php/gola/92-studio-prof-a-tanzariello/laringe/esami/724-valutazione-percettiva-della-voce (cit. on p. 7).

[8] *Software Instruction Manual "Multi-Dimensional Voice Program (MDVP)" Model 5105*. KayPENTAX. Chap. Appendix C, pp. 135–189 (cit. on p. 14).

[9] Robert B. Randall. «A history of cepstrum analysis and its application to mechanical problems». In: *Mechanical Systems and Signal Processing* 97 (2017). Special Issue on Surveillance, pp. 3–19. ISSN: 0888-3270. DOI: https://doi.org/10.1016/j.ymssp.2016.12.026. URL: https://www.sciencedirect.com/science/article/pii/S0888327016305556 (cit. on p. 15).

[10] Alan V Oppenheim and Ronald W Schafer. «From frequency to quefrency: A history of the cepstrum». In: *IEEE signal processing Magazine* 21.5 (2004), pp. 95–106 (cit. on p. 15).

[11] Tech. rep. available online. URL: https://realpython.com/logistic-regression-python/ (cit. on p. 31).

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