POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering,
Biomedical Intrumentation

Master's Degree Thesis

Extraction and selection of vocal features for the assessment of surgeries and rehabilitation of post laryngectomy patients

Supervisors

Prof. Alessio CARULLO

Prof. Alberto VALLAN

Candidate

Giulia RESIO

December/2022

Abstract

Open Partial Horizontal Laryngectomies (OPHL) are diffuse surgeries for laryngeal carcinomas, leading to post-intervention complications in the execution of primary activities, as phonatory abilities. In worst cases, the surgery encompasses the removal of both vocal cords (type II, III OPHL) and the outcome is a very hoarse and breathy voice, named "substitution voice". Patients must follow a rehabilitation path to partially restore the abilities impaired by the surgical procedures, and auditory perceptual evaluation scales as the INFVo are commonly used in the clinical field to assess the effectiveness of the rehabilitation on the voice quality. The goal of this work is to define a procedure based on voice analysis of patients, by extracting representative parameters and providing objective data on rehabilitation results. The data set used in this thesis was supplied by San Giovanni Bosco Hospital (Turin) and consists of 85 patients divided among the type of operations they underwent: 22 for OPHL-I, 32 for OPHL-II, and 31 for OPHL-III. All the acquisitions were made with an in-air microphone system and include vocalization of the sustained vowel /a/ and a phonetically balanced speech for each patient. First, signals were pre-processed with the software Audacity and Matlab (R2022a); then parameters were extracted for the whole recordings, but only those related to harmonic frames were considered for feature extraction. The harmonic frames were selected using two different criteria: the first one is based on the Harmonic-to-Noise Ratio (HNR) and the second one on the Spectral Kurtosis (SK). Examples of extracted features are SK, HNR and fundamental frequency (f_0), and other parameters in the spectral and cepstral domains, such as Cepstral Peak Prominence Smoothed (CPPS) and Mel-Frequency Cepstral Coefficients (MFCC). Each parameter is represented as a probability distribution, through descriptive statistics (indices of central tendency and range, measures of variability). To these, nine parameters were added for the vowel /a/ to evaluate period and amplitude stability, resulting in 198 features for the vowel /a/ and 189 for the balanced speech. Data were classified by the Logistic Regression (LR) model, by comparing first the type of intervention (OPHL I vs OPHL II, III) and then the patients within the worst cases (OPHL II, III) by dividing them into two classes based on index I (intelligibility) of the INFVo

scale. Feature selection relied on the accuracy (Acc) or Area Under The Curve (in case of a tie) of the LR model, trained using a single feature and then a combination of 2,3,4 features with low ($R^2 < 0.5$) and statistically significant (p-value< 0.05) correlation. Eventually, a method was proposed to quantify the role of the expanded uncertainty $U(p)$ of the probability p provided by the LR model, considering variances and covariances of model parameters; confidence intervals were created for each probability, thus the "non-classified" class was introduced, to be excluded in Accuracy evaluations. New metrics as Fraction Of Classified (Foc) and Realistic Accuracy (Acc_{real}) were proposed to test classification performances. Classification gave good results, mainly by SK method, balanced speech, OPHL I vs II,III with Acc values up to 96.5%, selecting Spectral Entropy (95-th percentile), f_0 (5,95-th percentile). New metrics were effective. For instance, a case of HNR method, balanced speech, OPHL I vs II,III selecting f_0 (range), HNR (skewness): Acc=94.1%, Acc_{real}=95.9%, Foc=0.87.

Acknowledgements

I miei sinceri ringraziamenti vanno
Al mio relatore Alessio Carullo, per avermi guidata nello svolgimento e nella stesura di questa tesi
Alla mia famiglia, a chi resta ed a chi è volato via
Alle mie amiche ed ai miei amici, mai distanti più di un passo

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter will provide an overview of the phonatory system and the mechanisms involved in the production of the human voice. There will be also an introduction to the major diseases affecting the larynx and to surgical techniques to treat them. Rehabilitative technics and classical methods used in clinical practice to evaluate the assessment of voice quality after rehabilitation will also be discussed, to better understand the new approach based on acoustic analysis on which this thesis work is based. Eventually, parameters related to the cepstral domain will be introduced.

1.1 Anatomy of the Phonatory system

The production of the human voice is the result of the cooperation between the Respiratory system, the Phonatory system, and the Resonatory system. The lungs are the main organs of the Respiratory system, they can be seen as the fuel behind voice production since they are responsible for the airflow that enables breathing. The Resonatory system includes the vocal tract from the trachea to the mouth, and it is fundamental in shaping the tone of a voice; all the organs in the oral cavity (tongue, teeth, lips, palate) are responsible for the generation of the consonants, by stopping the flux of air coming from the lungs. The nasal cavity has an active role as well in the production of the human voice [1]. The key element of the Phonatory system is the larynx, also known as the "voice box" and it is located in the anterior neck, between the pharynx and the underlying trachea,

and anterior to the esophagus (fig. 1.1). The larynx is a mucous membrane formed

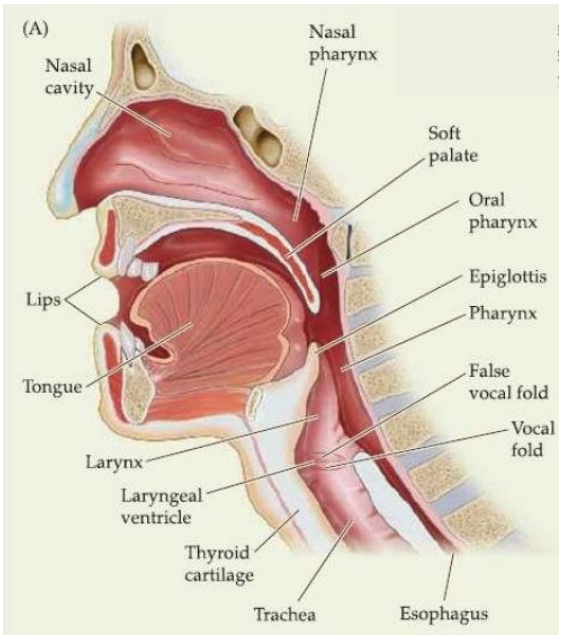


Figure 1.1: Anatomical position of the larynx in the neck [2].

by nine cartilages (figure 1.2), three unpaired (*thyroid, cricoid, epiglottis*), and six paired (*arytenoid, corniculate, cuneiform*). The *thyroid cartilage* is the largest cartilage and functions as a protective shield. The *cricoid cartilage* is basically a ring that encircles the trachea and is found in the inferior part of the larynx. The *epiglottis* is an elastic flap, that allows the passage of air into the larynx, trachea, and lungs. The epiglottis plays an important role in the protection of the lower respiratory tract; as a matter of fact, it lets the air in towards the trachea during breathing, but it closes while swallowing, to block food and drinks from going down into the trachea [2][3]. The larynx is divided into three sections: the *supraglottis*, which goes from the epiglottis down to the ventricular folds (false vocal cords), the *glottis*, which contains the true vocal cords and the *subglottis*, that goes down to the cricoid cartilage (beginning of the trachea). The vocal cords (or vocal folds) are located within the larynx at the top of the trachea and they are part of the glottis, a portion of the laryngeal cavity formed by the vocal folds and the *rima glottidis*, an opening between them. There are actually four vocal cords, but only two are directly involved in the phonation process. Indeed, above both sides of

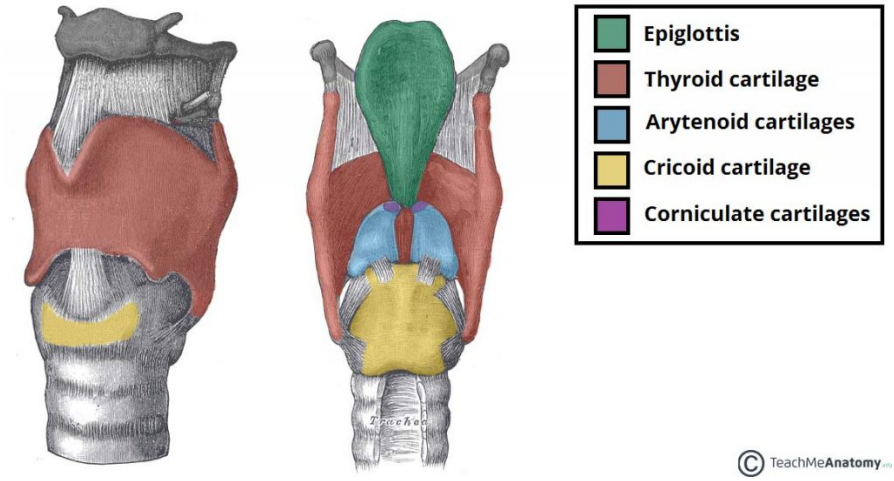


Figure 1.2: Arrangement of cartilages within the larynx [3].

the glottis there are two ventricular folds, also referred to as false vocal cords, and used to produce deep sonorous tones. As evidence, the true vocal cords appear to be thicker, wrapped in muscular fibers, and with a narrow gap between them, while the ventricular folds look thinner (figure 1.3). The phonation cycle is divided into the opening phase when the vocal cords are separated (inspiration, figure 1.3) and the closing phase, when the space between the vocal cords is slightly reduced (exhalation). The airflow produced by the lungs creates pressure below the glottis, which increases as the vocal cords are in complete adduction, during the closing phase. The vocal cords remain closed until the subglottal pressure gets high enough to push them apart, producing a negative intraglottical pressure, which pulls the vocal folds back and closes the glottis. The cycle is then repeated and it allows sustained vibration of the vocal folds, by creating an acoustic wave that propagates through the vocal tract between the trachea and the mouth [4]. Human sounds can be divided into two categories, according to the way they are produced: voiced sounds came from the vibration of the vocal folds, and originate during exhalation, typical examples are vowels /a/ and /i/; unvoiced sounds, also called "voiceless consonants" (for instance f, k, p, t, and s in the Italian language), on the opposite, originate in the opening phase of the phonation cycle: the airflow goes from the lungs to the mouth, where the tongue, teeth, and lips engage to modulate the sounds [5].

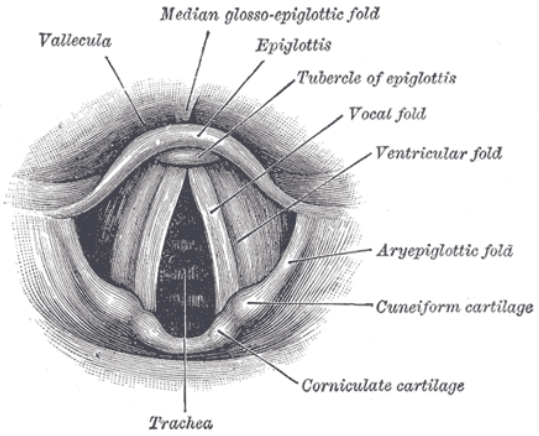
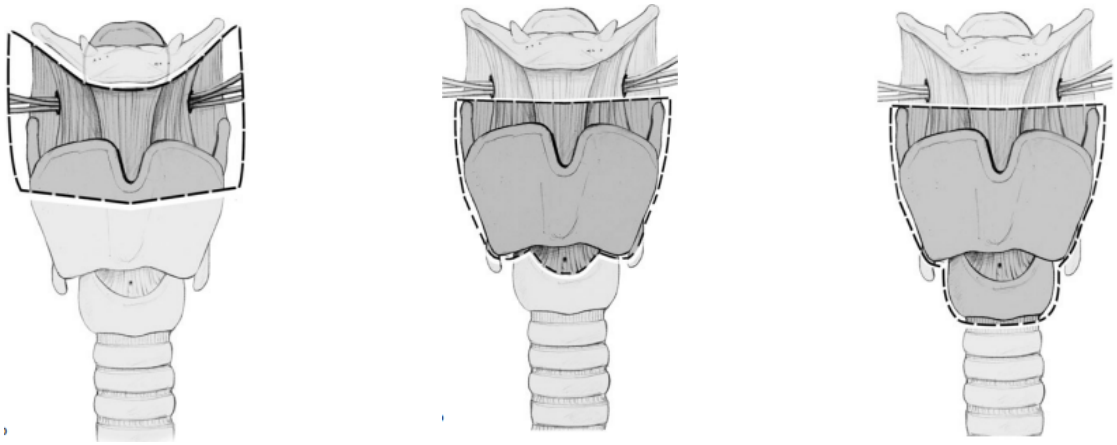


Figure 1.3: Laryngoscopic view of the interior of the larynx during the opening phase of the vocal folds. [6]

1.2 Diseases and surgeries

Excessive use of the vocal folds can cause laryngeal trauma and lead to laryngitis. Laryngitis is an inflammation of the larynx and may be chronic or acute (in this case is related to a viral upper respiratory tract infection). It can cause dysphonia, leading to hoarseness, pain, and coughing [7]. In worst cases, dysphonia, dysphagia, and dyspnea can be symptoms of laryngeal cancers. The carcinoma is usually found after a visual inspection with a laryngoscopy, which is possibly followed by a biopsy to confirm the diagnosis. The main risk factors include smoking and alcohol abuse [8]. Laryngeal cancers have been treated throughout the years both with radiotherapy and surgeries. The total excision of the larynx was a very common practice back in the years, but it led to complications in swallowing and in the quality of the voice. From the 90s it became quite popular the use of radio-chemo therapies in substitution of the total exportation of the larynx or in cooperation with less invasive surgeries, known as subtotal laryngectomies. Laser cordectomy has also emerged in the last decade in the treatment of glottic carcinomas, but it suffers from more difficult reproducibility of results. Subtotal laryngectomies are very delicate surgical interventions that must be planned very carefully, evaluating variables related to the patient and the type of carcinoma involved in the treatment. To actuate a laryngectomy it is necessary to create an opening in the trachea, named tracheostoma. This allows the patient to breathe during the operation

and in the postoperative period, through a cannula. Meanwhile, the tracheostoma is permanent for total laryngectomies, it is not in subtotal ones [9]. One of the less-invasive operations is the Open Partial Horizontal Laryngectomy (OPHL) a useful tool in the management of radio-resistant laryngeal cancers. There are three



(a) OPHL-I (HSL) **(b)** OPHL-II (SCL) **(c)** OPHL-III (STL)

Figure 1.4: Different OPHL types [9].

types of OPHL (figure 1.4) which take their names after the lower limit of resection, and each of them is involved in the treatment of a specific laryngeal cancer [10]. Supraglottic carcinoma extends from the epiglottis to the top of the glottis and rarely attacks the thyroid cartilage and the vocal folds. Type I OPHL (supraglottic laryngectomy) is suitable for tumors in the upper section of the larynx; the surgery encompasses the removal of the whole supraglottis (epiglottis and upper half of the thyroid cartilage) but preserves the arytenoids and the true vocal cords, while the vestibular folds are removed. Glottic carcinomas go deeper down the larynx and may extend into the cryco-arytenoid cartilages, so more invasive surgeries are needed in their treatment. Type II OPHL (supracricoid laryngectomy) includes the resection of the entire thyroid cartilage, down to the upper edge of the cricoid ring. Type III OPHL (supratracheal laryngectomy) consists of the resection of the whole supraglottis, glottis (vocal folds included), and the cricoid ring. Depending on the location and the severity of the cancer, the surgeries listed above may change the amount of supraglottis excised, in particular, the upper part of the epiglottis may not be or be preserved. Both type II, III OPHL encompass the removal

of the vocal folds, and the patients have to resort to a new way of producing sounds through "substitution voices". Hence it is fundamental for the patient to be involved in a rehabilitation path during the postoperative period. The primary goal of this journey is to help the patient to restore his swallowing abilities. After decannulation, the patient has to learn to breathe again and the last aim is to restore the phonatory function [11].

1.3 Vocal tasks and perceptual scales

Voice quality is a definition that includes all the perceptual dimensions of the spectral envelope and its changes in time; when patients undergo invasive surgeries such as total laryngectomies or partial laryngectomies, they have to follow a rehabilitation path to partially restore the abilities impaired by the surgical intervention; the evaluation of the quality of voice is an important milestone to assess the effectiveness of the rehabilitation on the voice quality, or simply to investigate the presence of a certain dysphonia. Voice quality may be affected by the irregularity in the vibrations of the vocal cords, giving an auditory perception of roughness [12]. The main methods used by speech-language pathologists to evaluate the quality of voice after rehabilitation involve perceptual rating scales. The most common perceptual rating scales are:

- GRBAS scale (Global Roughness Breathiness Asthenia Strain): voice quality is rated along five parameters; each one is scored from 0 (normal voice) to 3 (pathological voice). G stands for the global hoarseness and gravity of the voice, R stands for the roughness, B stands for the extent of air during phonation, A stands for the lack of power in voicing, and S for the strain (hyperfunction). Recently was added parameter I to indicate the instability of the voice over time [13] [14].

- INFVo: this scale rates the parameters I (overall impression, intelligibility), N (unintended additive noise), F (fluency), and Vo (quality of voicing, sound voiced or unvoiced). All parameters are evaluated from 0 (good perception) to 10 (bad perception) [15] [16].

- MPT (Maximum Phonation Time): this is not a scale, but an index of time (usually s) that gives information about the efficiency of the respiratory mechanism during phonation. The patient is asked to take a deep breath and produce an /a/ vowel for as long as he can [17]. It is often used to measure the assessment of a voice since it is a fast and non-invasive method. Moreover, its value is expected to be related to the fluency parameter (F) of the INFVo scale: the longer MPT, the higher the fluency in the speech [16].

Perceptual scales are usually combined with self-assessment scales, which do not require the presence of a clinician but rely on the self-evaluation of the patient who underwent the surgical intervention. These scales suffer from the bias that comes from the self-evaluation of the patients; it is not uncommon for patients to perceive their vocal qualities as better than they turn out to be. Examples of auditory self-evaluation scales are:

- VHI (Voice Handicap Index): This test consists of 30 statements on voice-related aspects. Each statement is evaluated by the patient with a score that goes from 0 (total reliability to the issue) to 4 (no reliability to the issue) and spaces from emotional, and physical to functional issues. The amount of handicap perceived increases as the final score obtained increases. Specifically, 0-30 indicates a low level of handicap, 31-60 medium level, and 61-120 serious amount of handicap [18].

- SECEL (Self-Evaluation of Communication Experiences after Laryngectomy): it consists of 35 specific questions split into general, environment, and attitude, to asses the communication skills of laryngectomized patients. One example of a question might be "Do you have difficulty yelling or calling out to people?" and the subject has to answer the question with a number from 0 (never) to 3 (always). Higher scores are an index of perceived difficulty with communication after the surgery, so further rehabilitation may be needed. This questionnaire was originally made in English but further translated into multiple languages [19].

The use of these auditory-perceptual scales is usually associated with a visual judgment based on a laryngeal examination. Both methods can be affected by the

subjectivity of the clinician (or the patient, if self-evaluated) or also by the type of vocal task analyzed. Commonly the rated vocal tasks include a sustained vowel and a continuous speech; the continuous speech has temporal and spectral variations due to voice onsets, vocal pauses, fluctuations in the fundamental frequency (f_0), and other factors; the sustained vowels, instead are held relatively constants. These different vocal behaviors of the tasks might lead to perceived differences in the severity of dysphonia; therefore it is appropriate to associate these as well with an acoustic analysis, providing objective, qualitative and quantitative data [20]. Several authors have proposed a method for the evaluation of pathological voices and substitution voices, based on the extraction and analysis of certain parameters from the vocal tracks of the involved subjects. Patients are usually asked to vocalize a sustained vowel and to utter a continuous speech, which could be free speech or a text to be read; a number of parameters deemed significant for evaluating the effectiveness of rehabilitation are then extracted from the recorded traces. Examples of suitable parameters extracted from uttering of sustained vowels are the fundamental frequency (f_0) and the Harmonic-To-Noise-Ratio (HNR) or measures of perturbation over time and frequency (shimmer and jitter). As far as concerns speech tasks, the most common parameters are HNR, f_0, Soft Phonation Index (SPI), MPT, and the Spectral Tilt. Alongside these parameters from the spectral and time domains, there are also parameters from the cepstral domain [21].

1.4 The cepstrum and its main parameters

To explain the parameters in the cepstral domain, it is necessary to take a step back on the origin of the human voice. Voiced sounds originate from the vibrations of the vocal folds and are then propagated through the vocal tract. The vocal signal has two main components: one related to the glottal pulses and another related to the vocal tract (real speech signal). The glottal pulses carry out a high-pitch signal, that is very noisy and is assumed as quasi-periodic. This signal is then filtered in the vocal tract to create the real speech signal, which carries information about the timbre of the voice (formants) [4]. Within this domain, it is common to use terms that are anagrams of spectral domain terminology (spectrum becomes "cepstrum",

frequency becomes "quefrency", harmonic becomes "rhamonic"). The cepstrum is mathematically represented as the spectrum of a logarithmic spectrum of a time waveform; whereas a spectrum gives information about the energy at harmonically related frequencies, a cepstrum brings information about the regularity of the harmonic peaks. The cepstrum operates within the domain of quefrency (which is a measurement of time) and its peaks are called rhamonics. Rhamonics occur at the quefrency at which the original time waveform has the fundamental frequency; the greater the regularity of a signal, the higher and more defined the rhamonic peak appears [20]. In figure 1.5 there are two examples: the upper graphic shows the smoothed cepstrum of a healthy voice and the lower graphic contains the smoothed cepstrum for an unhealthy voice; the first rhamonic peak at 5.8 ms is of easy detection in the upper graph, while it is not easily detectable in the lower graph. The cepstrum can be used for pitch detection, specifically, the first rhamonic can be used to go back to the frequency domain and understand what is the fundamental frequency of the vocal signal. A parallelism can be found between rhamonics in quefrency domain and the Dirac delta in the frequency domain when applying Fourier transform on a sinusoidal wave.

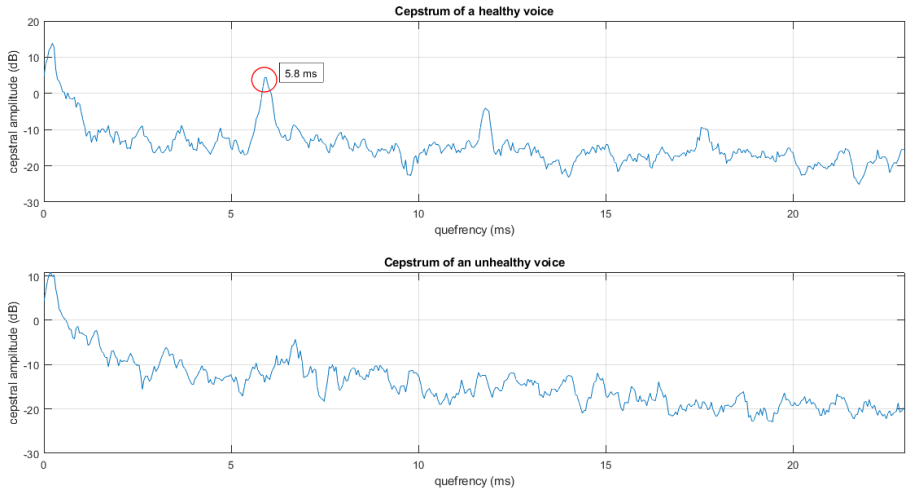


Figure 1.5: A smoothed cepstrum of a vocal signal of a healthy voice against the smoothed cepstrum of an unhealthy voice.

The speech signal x(t) can be mathematically described as the convolution in time

represented in eq. 1.1

$$x(t) = g(t) * v(t) * r(t) \tag{1.1}$$

where g(t) is the glottal component (modeled as an indefinitely long train of pulses), v(t) is the impulse response of the vocal tract and r(t) represents the effect of the acoustic wave radiation at the lips. The result is a quasi-period signal with a fundamental frequency equal to the one of the glottal signal. Fourier transformed is then performed and the convolution becomes a product, then the power spectrum is estimated as in eq.1.2

$$X(f)^2 = G(f)^2 \cdot H(f)^2 \tag{1.2}$$

where H(f) contains the contributes of the vocal tract and the acoustic wave radiation. By calculating the logarithm, the product is converted into the sum of two components

$$log(X(f)^2) = log(G(f)^2) + log(H(f)^2) \tag{1.3}$$

As a matter of fact, the spectrum of a vocal signal is a quasi-periodic signal multiplied by an envelope, that shows the slow variations of the signal and represents the signal in the vocal tract. The logarithmic trick comes to the rescue to separate the two components of the voice signal, which can be easily isolated with a filter. The most important information of the speech is contained in the envelope of the spectrum and this somehow explains why cepstrum is the result of a Fourier transform applied to a logarithmic spectrum. The filtering technique is called "liftering" and was introduced to emphasize the periodic components of the log spectrum and to enhance the detectability of echoes from a signal. The helpful information about the speech can be found in the lower end of the quefrency axis (below the first rhamonic peak) while the higher end (rhamonics peaks in general) carries out information about high-pitch noise (glottal pulses) [22] [23]. One of the most reliable cepstral measures of dysphonia severity is the CPP (Cepstral Peak Prominence). CPP was first introduced to assess the quality of breathy voices but then extended to the evaluation of overall voice quality. A variant was later proposed, called smoothed CPP (CPPs, figure 3.3) that provided higher correlation with breathiness, by adding smoothing operations both in temporal and cepstral domains. Both variants are expressed in dB and represent the difference between the most prominent peak of the smoothed cepstrum (first rhamonic) and the value

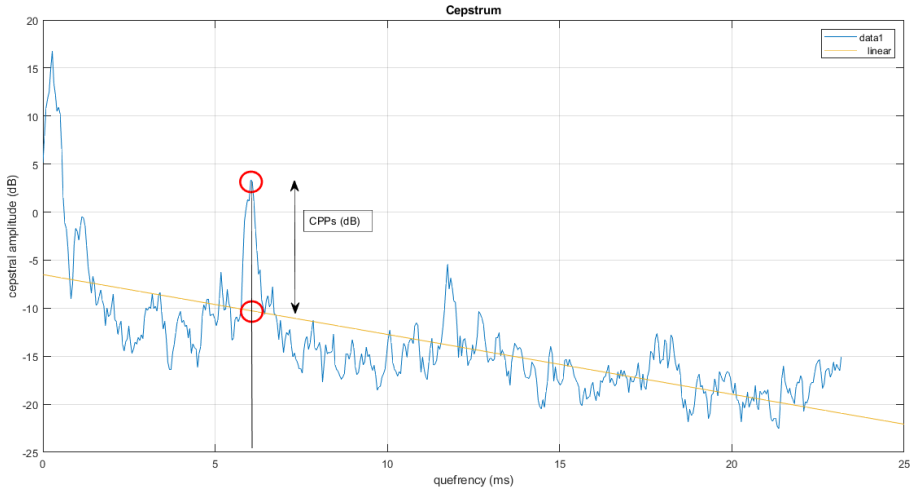


Figure 1.6: Graphic representation of CPPs of a random subject with a healthy voice.

at the same quefrency on the line of regression that correlates the quefrency and the amplitude of the cepstrum. CPPs appears to be meaningful in voiced signals rather than unvoiced signals, due to the direct relationship between their amplitude and the fundamental frequency. Furthermore, has been found evidence of an inverse relationship between the cepstral peak and perturbation measures, such as jitter and shimmer [23].

MFCCs (Mel Frequency Cepstral Coefficients) come from the cepstral domain as well. They are obtained through a series of steps applied to the original signal, and they were developed with the purpose of developing a system as close to model human hearing as possible. This model is based on a bank of triangular filters that will be better explained in the following chapter. MFFC have often been used in speech recognition systems and even in the clinical field, in the detection of Parkinson's disease [24]. MFCCs seem to be useful parameters in the evaluation of substitution voices of patients subjected to OPHL as well, alongside other spectral parameters (Spectral kurtosis and Spectral Entropy) [21].

Chapter 2

Materials and methods

This work has been the prosecution of a work already done in a previous master's thesis, which analyzed the phonetically balanced speech of 32 subjects that underwent type II, and III OPHL. In this study, the data set has been extended to all kinds of OPHL (type I, II, III) and both for phonetically balanced speech and sustained vowel /a/. The following chapter is set to describe the methods used, following the guidelines of voice analysis. The used data sets will be presented alongside the voice parameters considered appropriate. The two methodologies (HNR, Spectral Kurtosis) used to distinguish harmonic from unharmonic frames will also be explained in detail. There will then be a theoretical introduction to the Logistic Regression (LR) model used for classification, and on the Kolmogorov-Smirnov Test. It will be described the method of feature selection based on the accuracy of the LR model and the validation method used in the Matlab (R2022a) environment, Classification Learner App. Eventually, the proposed method to investigate the role of the expanded uncertainty U(p) of the probability p provided by the LR model will be discussed.

2.1 Data

The used data set was provided by San Giovanni Bosco Hospital (Turin) and includes a total of 85 Italian subjects with an average age of 63 years and mostly male. Subjects are divided among the type of operations they underwent: 22 for

OPHL-I, 32 for OPHL-II, and 31 for OPHL-III. All the acquisitions were made with an in-air microphone system, with a resolution of 16-bit at a sample rate of 50 kSa/s, and include vocalization of the sustained vowel /a/ and the phonetically balanced speech "*Notturno*":

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

A phonetically balanced speech is a type of continuous speech in which the words have the phonemes occurring at the same frequency at which they occur in normal conversations in that specific language (Italian, in this case) [25]. All files are presented in *.wav* or *.mov* format. The main purpose of the study was to find some valid vocal parameters that could be representative of the vocal quality of substitution voices and thus give an objective evaluation of the rehabilitation course of post-laryngectomized patients. Type I OPHLs preserve the vocal cords, while type II, and III OPHLs remove them completely, hence the patients have to speak with "substitution voices". It turns out to be intuitive that a substitution voice could sound noisier and less harmonic than a traditional voice, and this can be ascertained more objectively by certain parameters. So the subjects were divided into categories to be distinguished, first by comparing the type of intervention (22 OPHL-I vs 63 OPHL-II, III) and then by dividing the patients within the worst cases into two categories based on index I (intelligibility) of the INFVo scale. INFVo scale values range from 0 to 10, increasing as vocal perception worsens [15]; I=5 was seen as a reasonable value to discriminate between good and bad voices. Three data sets (table 2.1) were created with this indicator: two with threshold I=5 (one balanced, the other not) and one (balanced) with threshold I=2.8, to separate the ones with the best vocal quality. Generally, as can be seen in table 2.1 the number of subjects categorized with a good quality of voice exceeds the worst ones. This trend continued in the distinction among the good ones, with I=2.8; in fact, in order to obtain a balanced data set, it was necessary to remove a random patient between the ones with an excellent vocal quality (according to I-index). The decision to use the I-index of the INFVo scale as a method to

Table 2.1: Summary of the three data sets obtained within patients who underwent type II, III OPHL (classification based on voice quality).

23 ($I \geq 5$) vs 40 (I<5, good quality of voice)
23 ($I \geq 5$) vs 23 (I<2.8, the best within the good ones)
23 ($I \geq 5$) vs 23 (I<5, random extraction within the good ones)

discriminate between good-quality voices and bad-quality voices was taken after a review of the literature [15] and a check of the information that came along with the recordings of the patients. This information involves personal data such as gender, age, profession, and values of different perceptual scales, including the INFVo. The

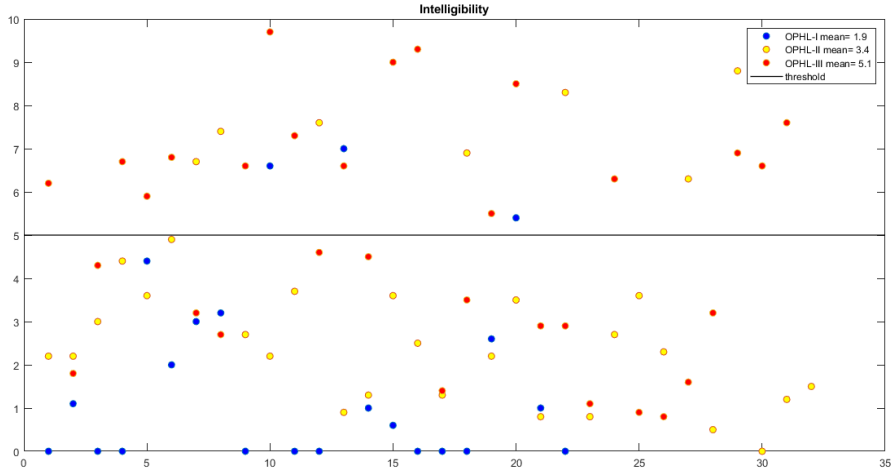


Figure 2.1: INFVo scale index I for patients submitted to OPHL-I (blue dots, I_{mean}= 1.9), OPHL-II (yellow dots, I_{mean}= 3.4), OPHL-III (red dots, I_{mean}= 5.2)

I values were collected and represented in a scatter plot (figure 2.1): blue dots refer to patients who underwent OPHL-I, yellow dots refer to OPHL-II and red dots refer to OPHL-III. It is noticeable at first sight how I values are almost all under the threshold at I=5 for OPHL-I, and more than the average have been given a 0, which means a very good perception of the voice. OPHL-II and OPHL-III have higher values, in particular, OPHL-III values are mostly above the threshold; this was not a surprise, since substitution voices are expected to appear less clear and defined when listening. The trend of index I among the patients confirmed the

reliability of the INFVo scale in the evaluation of substitution voices and therefore allowed the creation of the three data sets according to index I.

2.2 Pre-processing

The pre-processing phase is common to both tasks (sustained vowel /a/ and balanced speech). All the recordings were listened to and manipulated with the aid of the software Audacity, to remove the parts at the beginning and at the end of the tracks, that were not useful for the purpose of this study. After that, the files were further pre-processed in Matlab (R2022a). All files were loaded in the Matlab environment and re-sampled at 44.1 kHz, which is the typical sample rate in digital audio. Subsequently, a control on the mean value for each signal was performed. Whenever the mean value was higher than the 20% of the RMS value, it was removed. This step was followed by normalization with respect to amplitude: specifically, the signal was normalized to the absolute value of the maximum of the original signal. As the last thing, silences were removed from the signal with the use of a fixed threshold, equal to a half of the RMS value of the whole signal. A fixed window of 1024 samples was shifted over the signal, to verify if the samples within the window were above the set threshold. In case of a positive test result, the portion of the signal was considered valid and saved in a specific array, otherwise, it was discarded. The risk of using a fixed threshold resides in the fact that voices with low intensity, due to phonatory problems might be identified as silent and therefore canceled, by committing an error; anyways this technique proved to be effective when listening to the audio tracks. The signal was then divided into harmonic and unharmonic frames, so as to proceed with the feature extraction step only on the portion of the signal deemed valid. To effectuate this separation, two criteria were used. The first method relied on the Spectral Kurtosis (SK) parameter and the second on the Harmonic To Noise Ratio (HNR). The methods, which will be extensively described in the next chapter, involve considering the above-threshold portion of the signal as harmonic and the below-threshold portion as unharmonic, thus invalid.

2.3 Feature extraction

All signals, already unbundled from silences, were previously divided into blocks of 1024-sample each (23.2 ms with sr= 44.1 kHz) and smoothed with a Hanning-type window. The extraction of the parameters coming from the spectral, cepstral, and time domains was then carried out for every block of signal, for all signals (unless specified otherwise); anyways, only the parameters related to the harmonic frames of the signal were saved and taken into account for feature extraction. In this way, each parameter was seen as a probability distribution over time and was represented through nine descriptive statistics: mean, median, mode, 5-th percentile, and 95-th percentile as measures of central tendency; range, standard deviation, as indices of variability; and skewness, kurtosis as shape parameters. For both tasks, the distributions of twenty-one parameters of quality (table 2.2) represented with the aforementioned statistics were extracted, while for the sustained vowel /a/ task nine perturbation parameters were added (table 2.3), to evaluate signal stability over time and frequency. The following description and terminology of all the perturbation parameters and SPI refer to the one contained in the software instruction manual of MDVP, Model 5105 [26]. The mathematical formulas contained in the manual have also been used as a reference to implement the realization of the aforementioned parameters in Matlab (R2022a).

2.3.1 Quality parameters

Mel Coefficients, MFCCs (no 13)
Spectral Entropy, SEn
Spectral Kurtosis, SK
Logarithmic Spectral Tilt, STdB (dB)
Soft Phonation Index, SPI (dB)
Root Mean Square value, RMS
Smoothed Cepstral Peak Prominence, CPPs (dB)
Fundamental frequency, f_o (Hz)
Harmonic to Noise Ratio, HNR (dB)

Table 2.2: Quality parameters.

Quality parameters evaluate the quality of voice and include parameters from time, cepstral and spectral domains. Among these, the root mean square (RMS) value was also calculated for each of the 1024-sample frames, into which the signals unbundled from the silences were divided.

Mel coefficients (MFCCs)

These coefficients are representative of the vocal tract part of speech (with slow variation in frequency) and exclude the component related to glottic pulses; the aim is to compress the information about the vocal tract into numeric coefficients. MFCCs are calculated with the Mel scale, which connects each value of frequency (Hz) with a subjective pitch in the Mel scale, trying to emulate as close as possible the perception of the human ear. The relation between frequency and Mel domain can be approximately described with Eq.2.1

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.1}$$

roughly associating 1000 Hz to 1000 Mel. As far as concerns this work, it felt appropriate to consider only the first 13 MFCCs, as substitution voices usually do not have helpful components at high frequencies; so the scale was approached by a bank of 13 triangular band-pass filters (Fig.2.2) spaced up to 1 kHz, specifically the first centered at 133 Hz and the last centered at 1.7 kHz. In the Mel scale, the filters are narrow in the lower frequencies and widen toward the higher frequencies to achieve better resolution at low frequencies as in the human auditory system [27]. MFCCs are obtained through a defined procedure, whose steps may slightly

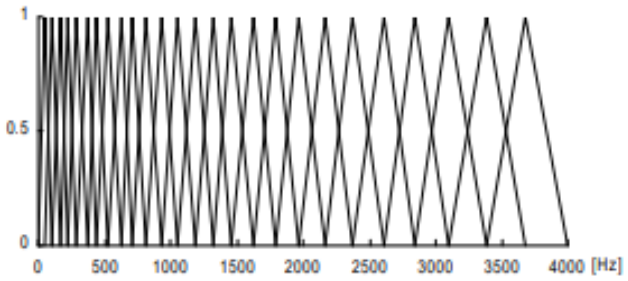


Figure 2.2: Frequency response of the bank of Mel filters [27].

vary according to the specific algorithm; here is reported a summary account of the necessary steps, after the review of a few algorithms [28][27]. The speech signal is first segmented into blocks of equal size, then each of the above frames is smoothed with a window (e.g. Hamming-type) and it is converted from the time domain to the frequency domain by implementing a FFT algorithm; to facilitate the FFT implementation, the signal is typically subdivided into frames of size equal to a power of two. Afterward, the magnitude frequency response is filtered by the bank of filters, to get a smooth magnitude spectrum. The last step before obtaining MFCCs is to move from the frequency domain to the quefrency domain and this is possible with the aid of DCT (Discrete Cosine Transform). Since the higher-order coefficients have a smaller amplitude than the first ones, an additional "liftering" (filtering in the quefrency domain) step may be needed, to rescale the coefficients and have comparable magnitudes between them. In Matlab R2022a there is the built-in function *mfcc* that performs these steps and returns the MFCCs. The function takes as basic input parameters the signal and its sample rate and gives as the default output 13 coefficients, which was appropriate for the case study. In this work, coefficients were calculated on signal portions of 1024 samples, previously windowed with Hanning-type windows and with 50% overlap; this process was repeated for the entire signal length, for all signals.

Spectral Kurtosis and Spectral Entropy

The AudioToolbox package in Matlab (2022a) provides built-in functions (*spectralKurtosis* and *spectralEntropy*) to calculate the spectral descriptors. Both functions have been implemented using Hanning windows of 1024 samples.

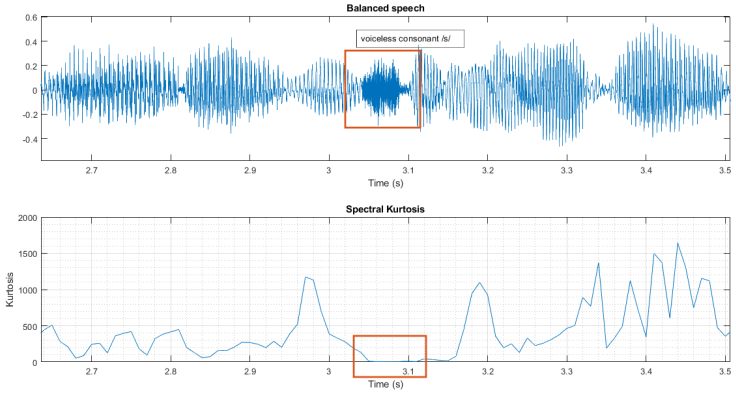
Spectral Kurtosis (SK): it measures the non-Gaussianity of the spectrum of a signal around its centroid; it is computed from the 4-th order moment, which can indicate the presence of series of transients and their positions in the frequency domain. When the background noise in the spectrum is reduced, the SK is likely to increase [21]:

$$SK = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^4 s_k}{(\mu_2^4) \sum_{k=b_1}^{b_2} s_k} \tag{2.2}$$

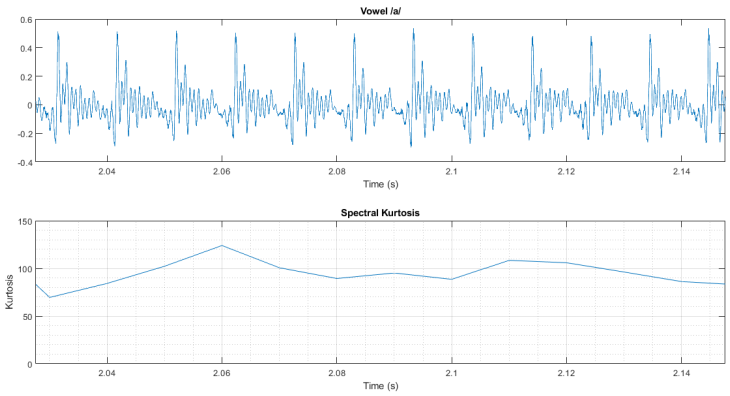
where

- b_1, b_2 are the bin limits of the interval analyzed (1024 samples in this case)
- f_k is the frequency at a specific bin
- μ_1, μ_2 are respectively the spectral centroid and the spectral spread
- s_k is the spectral value at a specific bin

In figure 2.3 are portrayed a few seconds of two vocal signals and their SK in that range of time, from patients having undergone OPHL type I and with little damage to their phonatory abilities. Figure 2.3a represents a short time portion of a balanced speech vocal signal, leading to an alternance between harmonic and noisy frames. As expected, SK is close to zero within the noisy portion of the signal (vocalization of the voiceless consonant /s/) and its values increase as the noise is reduced, reaching values greater than three orders of magnitude. On the opposite, for the sustained vowel /a/ case in figure 2.3b, it leaps to the eye the harmoniousness of the vocal signal in the time domain, mainly composed by harmonic frames, and SK shows values all above 50.



(a) Balanced speech, in red the pronunciation of consonant /s/

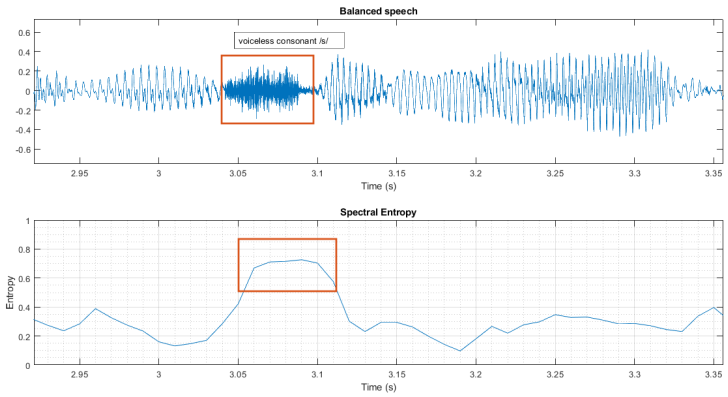


(b) Sustained vowel /a/

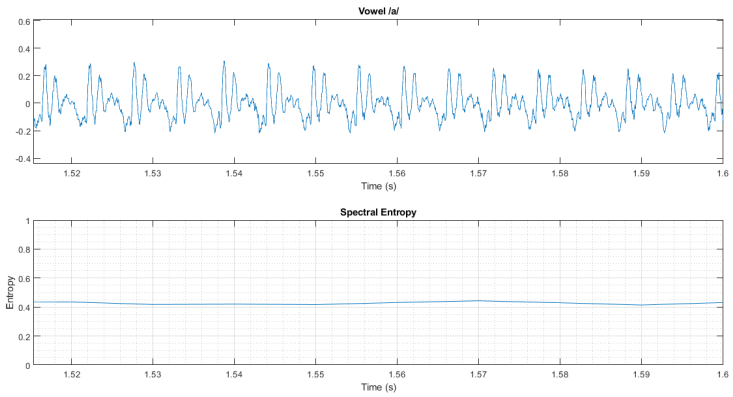
Figure 2.3: Spectral Kurtosis (calculated on the whole signal, unbundled from silences) of two random OPHL-I patients, in the cases of balanced speech and vowel /a/.

Spectral Entropy (SEn): is the Shannon entropy calculated on the normalized power distribution of the signal in the frequency domain. It is largely used in speech recognition to distinguish between voiced and unvoiced since regions of voiced speech have lower entropy compared to regions of unvoiced speech:

$$SEn = \frac{-\sum_{k=b_1}^{b_2} s_k \cdot log(s_k)}{log(b_2 - b_1)} \tag{2.3}$$



(a) Balanced speech, in red the pronunciation of consonant /s/



(b) Sustained vowel /a/

Figure 2.4: Spectral Entropy (calculated on the whole signal, unbundled from silences) of two random OPHL-I patients, in the cases of balanced speech and vowel /a/.

where

- b_1, b_2 are the bin limits of the interval analyzed (1024 samples in this case)
- f_k is the frequency at a specific bin
- s_k is the spectral value at a specific bin

In figure 2.4 are reported examples of SEn evaluation over time either for a balanced speech case, either for a sustained /a/ vowel case of a random OPHL-I patient, with low vocal impairments. As expected, the chart associated with the balanced

speech (fig 2.4a) has SEn values varying according to the pattern of the signal over time, having its peak during the utterance of the voiceless consonant /s/.
In the sustained vowel case instead, (fig 2.4b) SEn sways among the same low value for all the duration of the signal (in the figure is displayed a small portion of time) and has no great variations; this due to the fact that the signal is composed by a harmonic pattern that is repeated over time.

Logaritmic Spectral Tilt and Soft Phonation Index

For both parameters the Fast Fourier Transform (FFT) was derived in specific intervals, then the RMS value for each interval was taken as the Energy. Each parameter is represented as the logarithm in base ten of the ratio among energy at low and high frequencies.

Spectral tilt: it shows a comparison between energy at low frequencies E_1= (60 Hz $\div$ 1 kHz) and high frequencies E_2= (1 $\div$ 5) kHz in the logarithmic scale.

$$St_{dB} = 20log\left(\frac{E_1}{E_2}\right) \tag{2.4}$$

Soft Phonation Index: it is the average ratio of the lower frequency E_{1spi}= (60 $\div$ 160) Hz harmonic energy to the higher frequency E_{2spi}= (1.6 $\div$ 4.5) kHz computed in dB. It is an index of how strongly vocal folds adduct during phonation; an increase in value is an index of incomplete or absent adduction.

$$SPI = 20log\left(\frac{E_{1spi}}{E_{2spi}}\right) \tag{2.5}$$

CPPs

It comes from the cepstral domain and it is the absolute difference between the rahmonic peak of a cepstrum and the value on the regression line at the same quefrency. The CPPs was extracted through a Matlab script, based on the definition of cepstrum described in section 1.4. The routine takes as input the original signal, which is firstly down-sampled at 22.05 kHz, and then CPPs is estimated on frames of 2 ms using a 1024 Hanning-type analysis window (of duration 46 ms). For each window, FFT is computed twice: the first time to obtain the spectrum from the time domain and the second time it is performed on the logarithmic spectrum to obtain the cepstrum. Thereafter the cepstra of each analysis window are smoothed in time (using a window of 14 ms) and in quefrency (using a seven bin averaging window). As the next-to-last step, a regression line has been calculated between 1 ms and the maximum quefrency value. In fact, the quefrency at the cepstral peaks corresponds to the inverse of the fundamental frequency that happens to be in the range of (60 $\div$ 300) Hz and 1 ms might be set as the imaginary threshold that

divides contributes of the spectral envelope of the spectrum (below the peak) from periodic contributes (above the peak). Finally, the CPPs is evaluated in dB as the difference between the cepstral peak and the value on the regression line at the same quefrency [29].

HNR (Harmonic to Noise Ratio) and f_o (fundamental frequency)

Harmonic to noise ratio is an index of the harmoniousness of the voice signal, and it is evaluated in dB. In this work, the HNR was computed in the time domain of the signal, with the method of auto-correlation (AC) [30].

$$AC = \int x(t)x(t+\tau)\ dt \tag{2.6}$$

Assuming the signal to be stationary (the statistical properties of the signal do not change in time) AC can be described with Eq. 2.6 as a function of delay τ.

$$HNR = 10\log_{10}\frac{AC_v(T)}{AC_v(0) - AC_v(T)} \tag{2.7}$$

Still assuming the signal to be periodic and corrupted by white noise, there is a global maximum in AC at zero lag (corresponding to the power of the signal) and a local maximum in AC at lag T (period of the signal). By normalizing the AC of the signal to the maximum AC value at zero lag, the relative power of the harmonic components at the numerator and the relative power of the noise at the denominator are obtained. In this way it is possible to rewrite the HNR formula in Eq.2.7 in terms of relative power, shown in Eq. 2.8 and taken as a reference to implement the Matlab routines to obtain the parameter.

$$HNR = 10\log_{10}\frac{AC_v(T)/AC_v(0)}{1 - AC_v(T)/AC_v(0)} \tag{2.8}$$

The AC maximum has to be searched within the time interval that includes the expected f_0 range. Fundamental frequency encounters variations according to gender and age; a range from 90 to 400 Hz was chosen to implement the Matlab algorithms. Two different algorithms were used for sustained vowel /a/ uttering and balanced speech, which differ from each other for the length of the signal

in which looking for the maximum AC. As for the balanced speech task, AC is calculated once for the whole selected portion of the signal of 23.2 ms and the search of AC maximum occurs here. As far as concerned the sustained vowel /a/ task was performed more accurate distinction within the pseudo-periods: initially the is an estimation of the length of the pseudo-period from the AC computed on the original signal, but then further research is performed. The signal is divided into several intervals shorter than the estimated period and, for each one, AC is calculated and the maximum value is researched. At the end of the process, there is a distinction between harmonic and unharmonic frames through control of the HNR value: whether HNR> -6dB the frame is considered harmonic, and HNR and f_0 estimations are taken into account, otherwise, they are discarded.

2.3.2 Perturbation parameters

Relative Jitter, Jitt (%)
Absolute Jitter, Jita (μs)
Relative Average Perturbation, RAP (%)
Pitch Period Perturbation Quotient, PPQ (%)
Shimmer Percent, Shim (%)
Shimmer, ShdB (dB)
Amplitude Perturbation Quotient, APQ (%)
Coefficient of Fundamental Frequency Variation, vf_0 (%)
Coefficient of Amplitude Variation, vAm (%)

Table 2.3: Perturbation parameters.

APQ, ShdB and Shim are equally measurements of shimmer, which shows the irregularity of the peak to peak amplitude of the voice signal. ShdB and Shim both evaluate the same type of amplitude perturbation, but with different measures.

APQ (Amplitude Perturbation Quotient)

It measures the irregularity of the peak-to-peak amplitude of the voice, by applying a smoothing factor of 11 periods. APQ increases in breathy and hoarse voices, it can be associated with the presence of a turbulent noise in the signal or with the

inability of the cords to support a periodic vibration.

$$APQ = \frac{\frac{1}{N-10}\sum_{i=1}^{N-10}\left|\frac{1}{11}\sum_{r=0}^{10}A^{(i+r)} - A^{(i+5)}\right|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.9}$$

ShbB (dB)

It evaluates the period-to-period variability of the peak-to-peak amplitude within a voice sample. It is an index of amplitude stability.

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20log\left(\frac{A^{(i+1)}}{A^{(i)}}\right)\right| \tag{2.10}$$

Shim (%)

It is the evaluation of the average absolute difference between the amplitudes of two consecutive periods, divided by the average amplitude:

$$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.11}$$

According to [30], adult voices with Shim values that are above 3.81% are considered pathological voices.

Jita, Jitt, RAP, PPQ are equally jitter parameters, used to quantify the frequency variation from cycle to cycle within the vocal signal (stability in frequency). Variations may occur as a consequence of the inability of the vocal cords to support a periodic vibration for a defined period and are usually typical of horse voices.

Jita (Absolute Jitter)

It is an absolute measure given in microseconds, thus its values are strongly related to the fundamental frequency of the voice signal. Feminine voices, (with higher f_o)

result in lower Jita values than male voices (lower f_o).

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

Jitt (Jitter Percent)

It is a relative measure (%) widely used in literature to analyze pitch perturbation. The fundamental frequency of a vocal signal has a low influence on the determination of the Jitt parameter:

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

According to [30] adult voices with a jitter percentage higher than 1.04 % are held to be considered pathological.

RAP (Relative Average Perturbation)

It is a relative measurement (%) and estimates the irregularity of the pitch period of the signal with a smoothing factor of 3 periods.

$$RAP = \frac{\frac{1}{N-2} \sum_{i=2}^{N-1} \left| \frac{T_0^{(i-1)} + T_0^{(i)} + T_0^{(i+1)}}{3} - T_0^{(i)} \right|}{\frac{1}{N} \sum_{i=1}^{N} T_0^{(i)}} \tag{2.14}$$

0.68% is usually set as a threshold to discriminate between healthy and pathological voices [30].

PPQ (Pitch Period Perturbation Quotient)

It is a relative evaluation (%) of the period-to-period variability of the pitch within the signal, with a smoothing factor of 5 periods. Hoarse and breathy voices may

have an increased PPQ.

$$PPQ = \frac{\frac{1}{N-4}\sum_{i=1}^{N-4}\left|\frac{1}{5}\sum_{r=0}^{4}T_0^{(i+r)} - T_0^{(i+2)}\right|}{\frac{1}{N}\sum_{i=1}^{N}T_0^{(i)}} \tag{2.15}$$

vAm and vf_o are relative parameters that reflect the variation of the fundamental frequency within vocal signals. The changes might be related to periodic or non-periodic frequency tremors, high jitter, or rising and falling pitch during the analysis.

vf_o (Coefficient of Fundamental Frequency Variation)

It reveals any variation of the fundamental frequency within the voice signal.

$$vf_0 = \frac{\sigma}{f_0} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}f_0^{(j)} - f_0^{(i)}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}f_0^{(i)}} \tag{2.16}$$

vAm (Coefficient of Amplitude Variation)

It is the relative standard deviation of the fundamental frequency and reflects the variation of f_o within the analyzed voice signal. The vAm increases either with random or regular variations of the signal.

$$vAm = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}A^{(j)} - A^{(i)}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \tag{2.17}$$

2.4 Spectral kurtosis method

According to [21] spectral kurtosis might be a good parameter to discriminate among harmonic and unharmonic frames of a vocal signal. As a matter of fact, it is related to the flatness of the spectrum around its centroid and its value decreases as the noise increases. The threshold value of 2 was the one implemented in the

previous thesis work, selected among the means of the statistical distributions evaluated for 10 healthy subjects, as the one with the lowest standard error (SE). By referring to figure 2.3 reported in section 2.3.1 it should be noted how the behavior of the SK varies according to the nature of the signal. As far as concerns the balanced speech in 2.3a the noisy portions of the signal, as uttering of the consonant /s/ have SK values way below the threshold at 2, almost always close to zero; the harmonic components instead shows SK values above the 2 threshold, occasionally even above 1000. On the opposite, for the sustained vowel /a/ case in figure 2.3b, it leaps to the eye the harmoniousness of the vocal signal in the time domain, mainly composed by harmonic frames, and SK shows values all above the set threshold at 2. This SK threshold method did a good job within the balanced speech task, whilst it encountered some issues in the vowel /a/ task. This might be related to the nature of the vocal signals: when vocalizing the vowel /a/ in normal vocal conditions, the output ought to be a train of harmonics or at least something that resembles it in the worst cases. Consequentially, the values of SK should be quite similar among them and above the threshold at 2, leading to complications in separating frames with this technique. During an early stage of the work, voiced and unvoiced parameters (Eq.2.18,2.19) for both tasks were also calculated [31].

- Percentage of phonation time:

$$Dt\% = 100 \cdot \frac{n_{voiced}}{n_{voiced} + n_{unvoiced}} \tag{2.18}$$

- Percentage of silence:

$$Dt\_v\_s\% = 100 \cdot \frac{n_{silence}}{n_{voiced} + n_{unvoiced} + n_{silence}} \tag{2.19}$$

An unusually high percentage of phonation time was found within the balanced speech task (almost all above 98%) and so it was decided to raise the SK threshold to 200, which is the mean of the distributions of the median from the healthy subjects in the previous work, to exclude more frame and obtain more realistic phonation time percentage values. The percentage of phonation time got lower than the previous outcomes, but this method was set aside since the results were too imprecise at the hearing. With a higher threshold, a great number of frames

was excluded from the harmonic ones, therefore classifying as unharmonic an excessive load of voiced sounds. So the chosen threshold at 2 was kept to perform classification, but Dt % and Dt_v_s % were precautionarily excluded from the classification parameters.

2.5 HNR method

HNR as previously described in Eq. 2.8 can be seen as the ratio between relative signal power and relative noise power and can be used as an index of the harmoniousness of the signal. Because the definition is written as a logarithm in base ten, to obtain negative HNR values it is necessary that the power related to the noisy component is higher than that related to the harmonic component of the signal. As far as concern a healthy voice, the harmonic component is expected to carry more power than the noisy one, which results in HNR$>$ 0 dB [32]. According to this assessment, first an attempt has been made with an HNR threshold at 0 dB to discriminate within harmonic and unharmonic portions of the signal. No troubles were encountered with the balanced speech case, and the method worked well, confirmed by listening to the vocal tracks. In contrast, problems were found in the case of the sustained vowel /a/, since an HNR value above 0 dB appeared to be too optimistic for patients who underwent OPHL-II and III; in fact, due to the severe damage of the phonatory system as a result of such invasive operations, the substitution voices could be affected more by hoarseness and roughness, and the level of harmonicity of the vocal signal could be seriously impaired. As a matter of fact, when the quality of voice of the patients was very bad, there was nearly no frame with an estimated HNR above 0 dB, so the threshold was lowered to -6 dB, so as not to lose the information of those patients. This fact can be checked by looking at the histograms of HNR distributions in Fig. 2.5; here are reported the distributions related to the harmonic components of the sustained vowel /a/ signal for all patients that underwent OPHL-I,II, III. The distributions for type II, and III OPHL show a high number of negative HNR, despite having lowered the threshold to -6 dB; however, this trend is not seen for the type I OPHL case, whose values are mostly positive. Moreover, the distribution for type I OPHL is quasi-symmetric, while in type II, and III OPHL cases are asymmetric, both with

positive skewness.

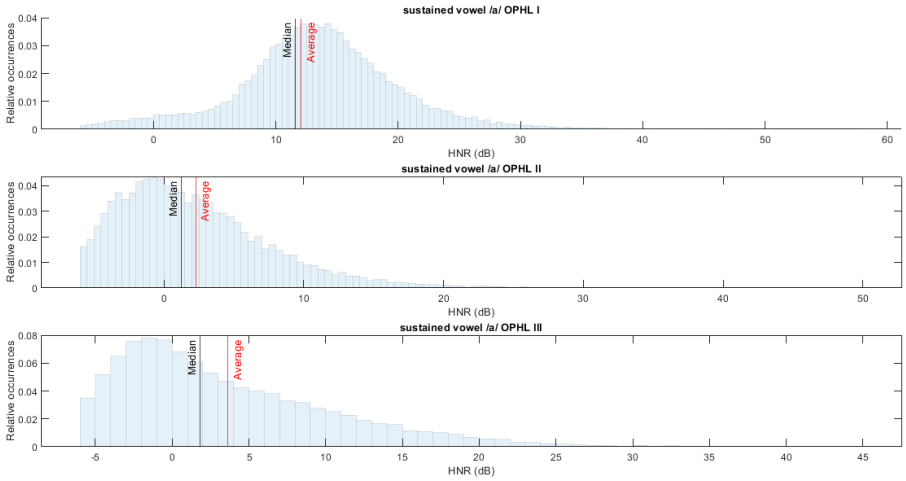


Figure 2.5: Histograms of the HNR distributions for the whole set of patients, who underwent type I, II,III OPHL.

2.6 Two-sample Kolmogorov-Smirnov test

This test was performed to research a discriminatory power among the features and to find features that might be representative of their own class; refer to the following Chapter, section 3.1 to look at the obtained results. The two-sample Kolmogorov-Smirnov test is a non-parametric hypothesis test that evaluates the differences among the cumulative distribution functions (cdfs) of the two sample data vectors over the range of values in each data set. The two-sided test uses the maximum absolute difference between the cdfs of the distributions of the two data arrays; the one-sided test uses the actual value of the difference, rather than the absolute value [33].

$$\begin{aligned} D* &= max_x(|\hat{F}_1(x) - \hat{F}_2(x)|) \quad \text{two-sided} \\ D* &= max_x(\hat{F}_1(x) - \hat{F}_2(x)) \quad \text{one-sided} \end{aligned} \tag{2.20}$$

The test was executed with the Matlab (R2022a) *ktest2* built-in function, which receives as an input the two arrays x1 and x2 and returns a decision made for the

null-hypothesis (H0) test; as H0 was taken that two samples belong to the same population, at 5% significance level (p-value < 0.05). The built-in function returns h=0 if the null hypothesis is not rejected and h=1 if it is rejected. The test was performed within each parameter among the subjects either belonging to the same class or different classes, excluding null values. During intra-class comparisons, the patients were compared two by two at a time, considering all the combinations obtained with Eq.2.27. In this case, all times when the test yielded h=0 were noted, to search for features that might be representative of their own class. As for inter-class comparison, all subjects in a class were compared to each patient belonging to the other class, resulting in a higher number of comparisons; the total number of tests is obtained by multiplying the number of subjects in both classes. In the OPHL-I (22) vs II, III (63) case it resulted in 1386 tests; in the OPHL-II, III cases (divided with I-index) resulted in 920 tests for the unbalanced data-set 920 tests and 529 for the balanced ones (refer to table 2.1 to see the data-sets). In this case, was noted when the test yielded h=1; in this way, it was possible to find out whether there were features with a particular discriminatory power that would allow the two classes to be distinguished.

2.7 Logistic Regression

Logistic regression (LR) is a non-linear statistical model, belonging to the class of Generalised Linear Models, that uses a logistic function (link function) to model a binary dependent variable. The logarithm of the odds in Eq.2.21 is a linear combination of independent variables X_i (predictors), and regression coefficients (β_i), where β_0 represents the intercept.

$$\begin{aligned} \log(\frac{p}{1-p}) &= \Theta^T x \\ \Theta^T x &= \beta_0 + \beta_1 x_1 + \beta_i x_2 + ... + \beta_N x_n \end{aligned} \tag{2.21}$$

The probability p returned by the model is a continuous function that can be described with the sigmoid function in Eq.2.22, obtained by inverting Eq.2.21. In binary classification problems, the probability can be seen as an ideal step function between 0 and 1, and the purpose of the LR model is to reduce the distance between

the sigmoid functions and the ideal step function [34]. This can be done by solving a minimization problem on the log-likelihood, by using deterministic or stochastic approaches, such as the least square difference, gradient descent, and the Newton method. The search for the best-regressed coefficients gives a set of best estimates of the coefficients and also an estimate of the coefficient variances and covariances which can be used to evaluate the goodness of the model.

$$p = \frac{e^{\Theta^T \cdot x}}{1 + e^{(\Theta^T \cdot x)}} = \frac{1}{1 + e^{-(\Theta^T \cdot x)}} \tag{2.22}$$

Since the probability returned by the LR model is a continuous function, to obtain a binary classification, it is compared to a threshold, typically set to 0.5. Whether $p <= 0.5$ the element is assigned to Class 0, else to Class 1. In this study, the negative Class 0 is associated with the healthiest possible condition in the data set analyzed and positive Class 1 with the worst possible condition (e.g in the data set OPHL-I vs II, III the healthiest condition is linked to OPHL-I that becomes the negative class, whilst OPHL-II, III becomes the positive class). The LR is part of a bigger family of learning algorithms, very popular in machine learning, called supervised learning algorithms; these have the characteristic to train the algorithm to make predictions with a data set of known features and responses. The

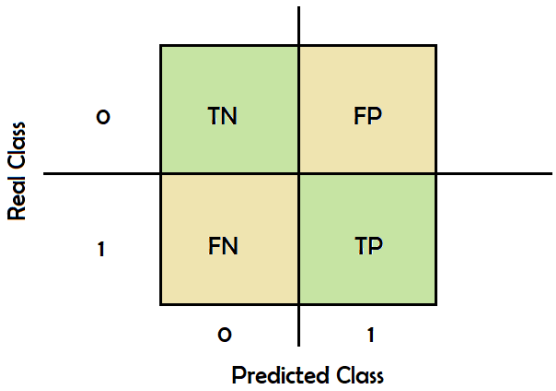


Figure 2.6: Example of a confusion matrix (0= Negative Class, 1= Positive Class)

probabilities returned by the LR model can be therefore compared with the known responses, to derive the quantities represented in a table named *confusion matrix* (Fig.2.6). True Positive (TP) is the total of subjects correctly classified as part of

the positive class 1, whilst True Negative (TN) is the number of subjects correctly classified in class 0. False Positive (FP) is the number of subjects incorrectly classified in class 1 and False Negative (FN) is the number of subjects incorrectly classified in class 0. These quantities are useful to evaluate metrics to assess the goodness of the classification. Common metrics are:

- Accuracy:
$$\frac{TP + TN}{TP + TN + FP + FN} \tag{2.23}$$
it is a measure of observational error, and it indicates how close the predictions are to their real value.

- Precision:
$$\frac{TP}{TP + FP} \tag{2.24}$$
it is a measure of observational error as well and indicates how close the predictions are to each other.

- Sensitivity:
$$\frac{TP}{TP + FN} \tag{2.25}$$
also called true positive rate (TPR), it measures the fraction of data belonging to the positive class. As far as concerns this work, a high TPR means the classifier has a good ability to correctly identify patients with worst voice conditions (Class 1).

- Specificity:
$$\frac{TN}{TN + FP} \tag{2.26}$$
also called true negative rate (TNR), it measures the fraction of data belonging to the negative class. According to this work, a high TNR reduces the likelihood that a patient with a healthy voice, will be classified as impaired; this allows the introduction of a metric called False Alarm = 1-Specificity.

- Area Under The Curve (AUC): it is the underlying area of the ROC curve, it can be estimated in the Matlab environment with the built-in function *perfcurve*. AUC values vary from 0, which is the worst measure of separability,

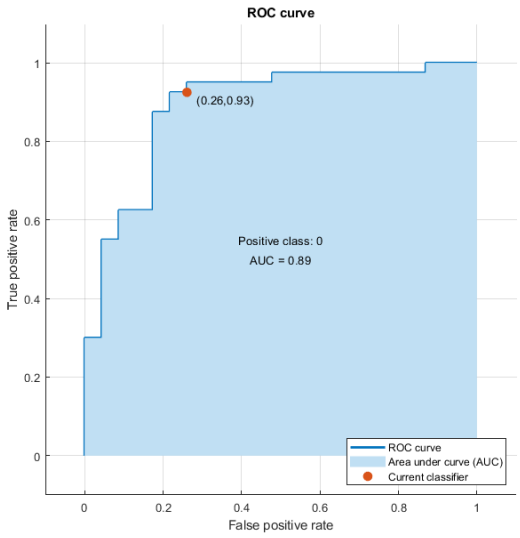


Figure 2.7: Example of a ROC curve.

to 1, which is the best measure of separability possible; when the AUC is 0.5, the model has no ability to distinguish classes. The ROC it is created by plotting the Sensitivity (TPR) against the False Alarm (FPR); an example is reported in Fig. 2.7

2.7.1 Feature selection

The feature selection (FS) algorithm was implemented in the Matlab (R2022a) environment and relied on the evaluation metrics computed after the training of a Logistic Regression model. The total amount of available features to train the LR model was 189 for the balanced speech and 198 for the sustained /a/ vowel task, but only the ones with low correlation and deemed to be statistically significant were taken into account. In fact, the FS algorithm was in charge of training the LR model with a single feature or a combination of 2, 3, or 4 features; but before proceeding, a check on the R^2 values of each pair of feature and their p-values was performed. Only features with $R^2 < 0.5$, which means a correlation R value in the interval $(0 \div 0.7)$, and p-value < 0.05 were considered valid for training. Once the training of the model was done, the FS indicated the more suitable feature (or combination of features) to validate the LR model, with 5-fold

cross-validation in the Classification Learner App, Matlab (R2022a). To select the feature (or combination of features) by which to validate the model, the FS algorithm compared the values of the accuracy extracted from the training case of the model and automatically indicated the feature (or combination of features) with the highest accuracy value.
In this thesis, the LR was implemented for binary classification with the built-in function *fitglm*. The function receives as input: the data to be classified, their real classes (0 and 1), and other settings, that allow the LR model to define the distribution of the response variable as binomial ('Distribution', 'binomial') and to set the logit function (Eq. 2.21) as the link function to utilize ('Link', 'logit'). The function returns as output the LR model, complete with probabilities and an estimate of the coefficient standard errors and covariances, useful to evaluate the goodness of the regression model. The implemented algorithm roughly follows a pattern that starts with the selection of the number of features k to combine and the creation of the combinations with the binomial coefficient in Eq. 2.27. The work was completed with the built-in Matlab (R2022a) function *nchoosek* that returns either the binomial coefficients or all the combinations; it receives as input the number k of features considered (1,2,3,4) and the number of total features n (198 for vowel /a/ and 189 for the balanced speech).

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2.27}$$

The number of chosen features k was set to range from 1 to 4 since an attempt was made with k=5 features, but the computational cost was so high that caused trouble in the Matlab environment. The number of the obtained combinations is listed in table 2.4 for both tasks. At a glance, it is noticeable how the number is very high in the case of k=4. The algorithm provides a double check on the features before performing LR, one based on correlation values and one based on their p-values. The coefficient of determination R^2 was easily calculated for each feature as the square value of the correlation, computed with the Matlab function *corr* that also returns the p-values of correlations. All features having a smaller p-value than the fixed threshold at 0.05 were assigned their R^2 values, otherwise, they were set to 1, to be further excluded in the following criterion based on R^2.

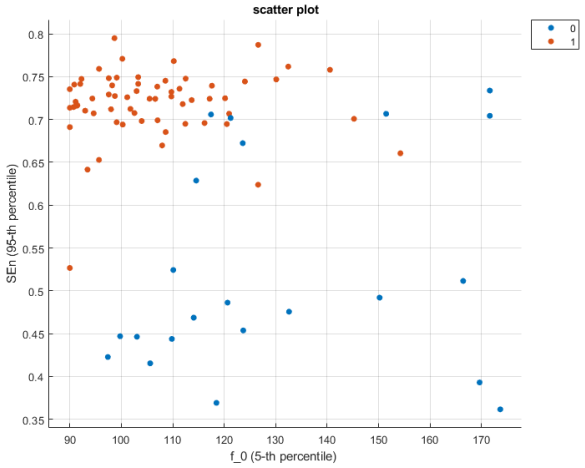
selected features	balanced speech	sustained vowel /a/
k=1	189	198
k=2	17766	19503
k=3	1107414	1274196
k=4	51494751	62117055

Table 2.4: Number of combinations for both tasks.

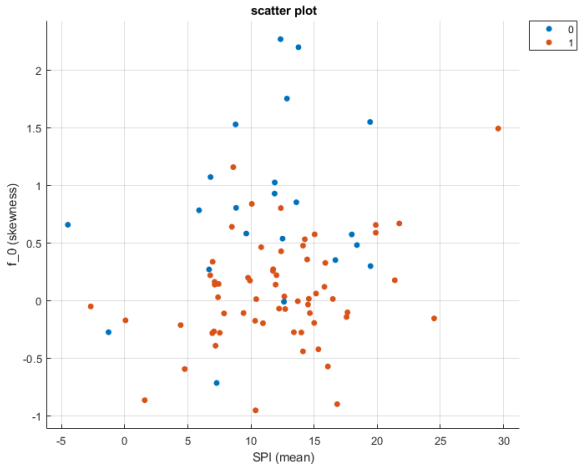
This second control was performed by taking the selected parameters two by two, a common practice in machine learning training: if the values of R^2 of the pair remained within the interval, LR was performed with the function *fitglm* with the set parameters explained before; else the combination of features was discarded and the algorithm moved to the next combination. The following step, whether all conditions were satisfied, was to carry out the predictions, by using the built-in Matlab function *predict*, which takes as input: the LR model, the matrix with the data to be classified, and other settings as 'Alpha', by which was set the condition on the maximum p-value to 0.05. This further condition on p-value was set to be sure to take into account only models with statistically significant coefficients. The probabilities values returned by *predict* were compared to the real responses to obtain the *confusion matrix*. At the end of the algorithm, with the values reported in the *confusion matrix*, the typical measures of classification for each combination were computed as well: Accuracy, Precision, Sensitivity, Specificity, and AUC. The algorithm automatically indicated a combination of features (for each k) with the best accuracy obtained in the training phase and this was used to validate the algorithm. Checking carefully, however, there were cases where multiple feature sets had the same maximum accuracy value. In this situation, validation was performed both with the combination of features returned by the algorithm with the highest accuracy and by selecting the combination of features with the highest AUC values among the ones with the highest accuracy.

2.7.2 Validation of the LR Model

The validation of the LR Model was carried out within the Classification Learner App, within the Matlab (R2022a) environment, using k-fold cross-validation, with k=5. This technique partitions data into five subsets of equal size; the process consists of using one subset among the five to validate the model trained with the other subsets; it is repeated five times so that each subset is used once for validation.
To perform the validation of the LR model, it was first necessary to load the matrix with the features and their classes in the Classification Learner App; after that, the features previously chosen with feature selection were manually selected, and the validation of the LR model was executed. The Classification Learner App returns the *confusion matrix* relative to the validation phase of the model and the values of the main metrics, such as the accuracy, and the ROC curves, with the relative AUC values; the App also gives the opportunity to recreate the Validated Model in the Matlab environment with the "Generate Function" button in the Export section.
Within the Classification Learner App, it is possible to get visual feedback on the features that are able to separate the two classes well, by means of a scatter plot comparing pairs of features. This was a useful tool to assess the goodness of the feature selection model proposed. Figure 2.8a is reported as an example of a pair of features selected in the prior feature selection phase, where it can be clearly seen the division between the two classes; on the opposite, in figure 2.8b there is a pair of features that demonstrates absolutely no ability to separate the two classes.
Cross-validation is a model assessment technique that works with any kind of supervised learning algorithms, so an attempt was made to use it as well with the Medium Tree model, using the features selected with the LR model; the method was then rejected, due to the poor performances obtained.



(a) f_0 (5-th percentile) against SEn (95-th percentile)



(b) SPI (mean) against f_0 (skewness)

Figure 2.8: Scatter plot for couples of features from type I vs II, III OPHL, balanced speech, SK method.

2.7.3 Expanded uncertainty for the LR model

As the last step in this work, a method was proposed, according to [34], to quantify the role of the expanded uncertainty $U(p)$ of the probability p provided by the LR model during the validation phase in the Classification Learner App, Matlab (R2022a). The LR-validated model was generated by the exported function from the Classification Learner App, which received as input the data set with features

and classes. The sensitivity coefficients were obtained by unwinding the partial derivatives of the probability function in Eq. 2.22 concerning the model coefficients (β) and are reported in equation 2.28:

$$\begin{aligned} \frac{\partial p_i}{\partial \beta_0} &= p_i \cdot (1 - p_i) \\ \frac{\partial p_i}{\partial \beta_j} &= F_j \cdot p_i \cdot (1 - p_i) \\ j &\in [1...N_F] \end{aligned} \tag{2.28}$$

where N_F is the number of considered features.
In this way, the standard uncertainty $u(p)$ was estimated through the uncertainty propagation formula, taking into account both variances and covariances of the model coefficients, which were all extracted from the validated LR model; each β coefficient is in fact associated with an uncertainty value (SE) and with a covariance value. The formula is reported in equation 2.29:

$$\begin{aligned} u(p) &= \sqrt{J_\beta \cdot COV_\beta \cdot J_\beta^T} \\ J_{i,j}(\beta) &= \frac{\partial p_i}{\partial \beta_i}; j \in [1...N_F + 1]; i \in [1...N_S] \end{aligned} \tag{2.29}$$

where N_F is the number of considered features, N_s is the number of samples in the data set and $J_{i,j}(\beta)$ is the Jacobian matrix of the model coefficients, COV is the variance-covariance matrix of the coefficients. The expanded uncertainty $U(p)$ was then obtained by multiplying $u(p)$ by a coverage factor of 2; this allowed the creation of intervals of confidence for each probability value returned by the LR-validated model. The expanded uncertainty was graphically represented by error bars, an example is given in figure 2.9.
According to the example in the figure, in general, the error bars of probabilities closer to 0 tend to narrow, while they tend to widen as they approach 0.5.
All subjects with probabilities ranges that intersected the set threshold at 0.5 were considered too doubtful to be involved in the binary classification, and the third class of "non-classified" was therefore introduced. For instance, subjects 2 and 20 in fig. 2.9 fall in this class. To have an objective perception of the effect of "non-classified" subjects on the overall classification performance, new metrics were introduced, such as the Realistic Accuracy (Acc_{real}) and Fraction

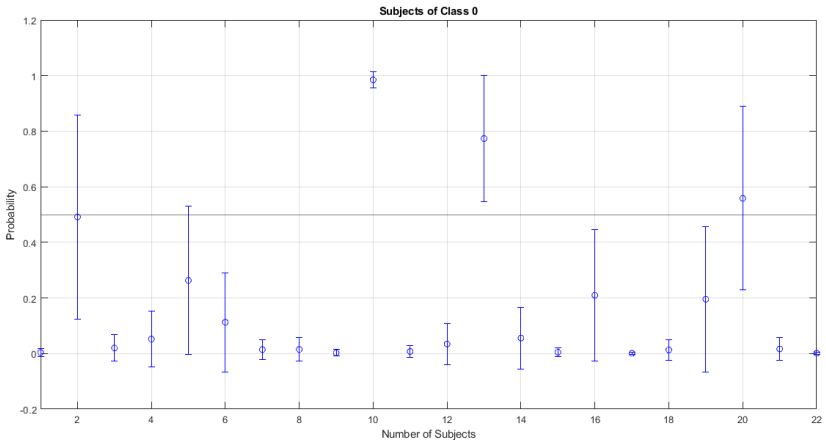


Figure 2.9: Examples of intervals of confidence for subjects of Class 0.

of Classified (FoC). The Realistic Accuracy is nothing but the calculation of the accuracy (as reported in eq. 2.23) by excluding the elements belonging to the "non-classified" class. The Fraction of Classified, on the other hand, is the ratio between the classified elements and the total number of elements, it is an index of the percentage of the subjects held to be classifiable. All results found during this analysis based on the expanded uncertainty of the probability returned by the LR-validated model are reported in the following chapter, section 3.3.

Chapter 3

Results

This chapter contains the most significant results found during this thesis work. All tables report the information in the columns according to the following legend: "Task" is the analyzed task, bs= balanced speech, sv= sustained vowel /a/; "Selected features" are the outcomes of feature selection, as explained in section 2.7.1; "Method" is the method used to discriminate among harmonic and unharmonic frames, with Spectral Kurtosis (SK) or Harmonic to Noise Ratio (HNR); the metrics as Area Under The Curve (AUC), Sensitivity (TPR), Specificity (TNR), Accuracy (Acc) refer to the training phase unless specified. To facilitate the reading of the reported data, refer to 2.1 which contains a summary of the data sets used during classification with the LR model and during the performances of the Kolmogorov-Smirnov test.

3.1 Kolmogorov-Smirnov results

The Kolmogorov-Smirnov test was performed with the distributions of the parameters, as a first attempt to look for features representative of a certain class. The test performances in some cases yielded interesting results, which were also confirmed in the classification with the LR model. Note to readings, the brown text in the following tables highlights a correspondence between the results of the Kolmogorov test and the ones from the feature selection, which relied on the training of the LR model with 1, 2, 3, and 4 features.

3.1.1 Inter-class comparisons

The most relevant results obtained during the performances of the inter-class Kolmogorov-Smirnov test are reported in this section. The total number of tests computed differs according to the data-sets numerosity and only the features with the highest amount of H0 rejected (test yielded h=1) are reported. Generally looking at the tabulated values, there are features that appear more often than others, thus indicating a strong discriminatory power. For what concerns the balanced speech cases, looking at the examples in tables 3.1, 3.2, recurrent features are the fundamental frequency f_0, the Mel Coefficients (MFCC5, MFCC7, MFCC8, MFCC9, MFCC12, MFCC6), and the Spectral Entropy (SEn). These insights of the test were also reflected in the results of feature selection of the balanced speech cases, reported in section 3.2.2, table 3.8 (for the type I vs II, III OPHL data set, SK method) in which the selected features happened to be statistics of SEn and f_0. Still, with regard to the case shown in table 3.1 the Kolmogorov-test rejected the null hypothesis 1385 times out of 1386 with the feature HNR, hence suggesting a strong discriminatory power of the feature; anyway, the classification result with the LR model did not confirm this insight.

Table 3.1: Kolmogorov-Smirnov test results for 22 OPHL-I vs 63 OPHL-II, III, SK method, balanced speech. Total number of tests: 1386.

H0 rejected	Feature	H0 rejected	Feature
1385	HNR	1381	SK
1384	f_0	1379	MFCC12
1381	SEn	1379	MFCC6

Table 3.2: Kolmogorov-Smirnov test results for OPHL-II,III, I<5 (40) vs $I \geq 5$ (23), SK method, balanced speech. Total number of tests: 920.

H0 rejected	Feature	H0 rejected	Feature
912	MFCC5	909	SEn
911	MFCC7	909	f_0
909	MFCC8	908	MFCC9

Figure 3.1 is an example of poor correlation between f_0 distributions, coming from pairs of patients belonging to the two different classes of the case described in table

3.1; each class has a particular distribution, which diverges from that of the other class in terms of central tendencies and variability as well.

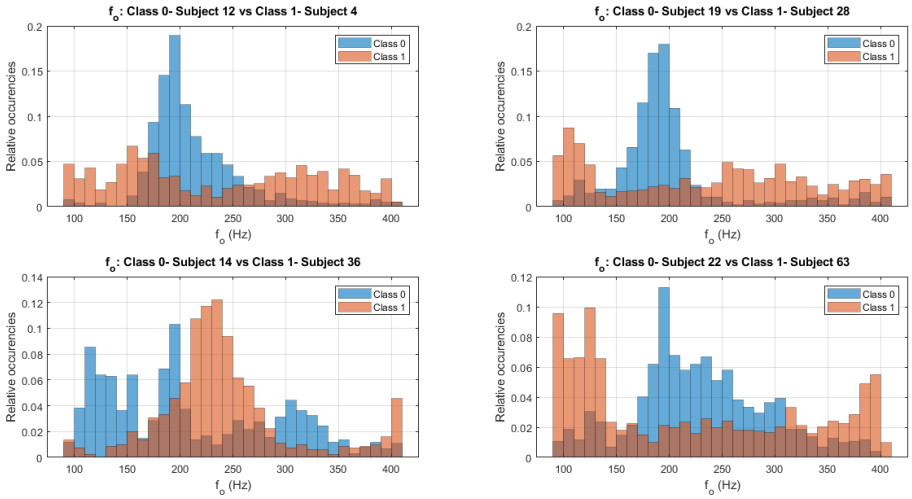


Figure 3.1: Examples of poor correlation among f_0 values for the balanced speech case reported in table 3.1.

As far as concerns the sustained vowel /a/ case (tables 3.3, 3.4), the fundamental frequency f_0 and the Mel Coefficients (MFCC5, MFCC7, MFCC8, MFCC9, MFCC3, MFCC11) pop up among the other parameters. The statistics of MFCC5 (mean) and MFCC9 (range) were two of the four parameters chosen (in the k=4 combination case) by the feature selection algorithm, as reported in the following section 3.2, table 3.9 (for the type II, III OPHL, 23 with $I \geq 5$ and 40 with I<5, HNR method). The skewness of MFCC5 was also selected in the k=2 combination case, reported in table 3.6 (for the type I vs II, III OPHL, HNR method). Figure 3.2, is given as an example of the different behavior of MFCC9 distributions between pairs of patients of two different classes, for the case reported in table 3.3. The distributions have a similar central tendency behavior but differ in terms of variability; for instance, the figure at the top right clearly shows that MFFC9 values are centered around 0 for the Class 1 subject while are distributed over multiple values in the case of the Class 0 subject.

Table 3.3: Kolmogorov-Smirnov test results for OPHL-II,III, I<5 (40) vs $I \geq 5$ (23), HNR method, vowel /a/. Total number of tests: 920.

H0 rejected	Feature	H0 rejected	Feature
903	f_0	895	MFCC3
902	MFCC7	894	MFCC5
896	MFCC9	891	MFCC8

Table 3.4: Kolmogorov-Smirnov test results for OPHL-I vs OPHL-II, III, HNR method, sustained vowel /a/. Total number of tests: 1386.

H0 rejected	Feature	H0 rejected	Feature
1385	f_0	1376	HNR
1378	MFCC11	1374	MFCC5
1376	MFCC3	1374	MFCC9

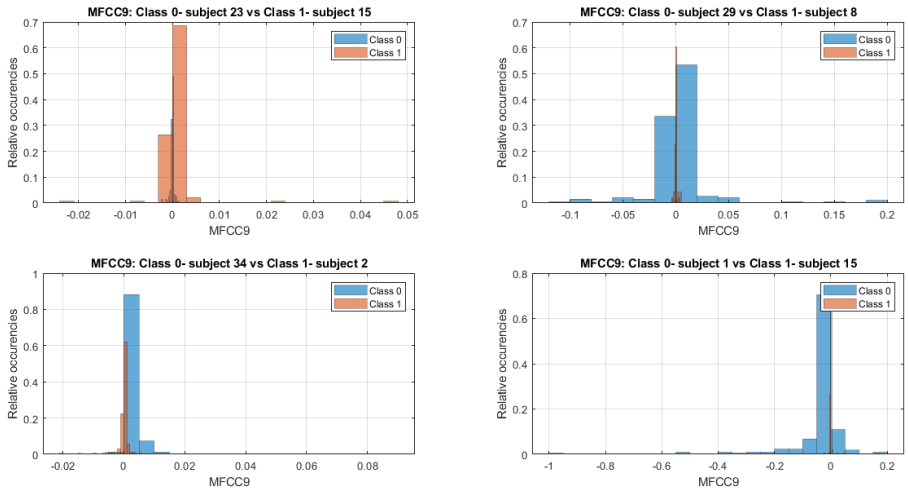


Figure 3.2: Examples of poor correlation among MFCC9 values for the case reported in table 3.3

3.1.2 Intra-class comparisons

As far as concerns the intra-class comparisons, the obtained results were not so meaningful. All cases in which the test did not reject H0 were noted (considering two samples belonging to the same class as H0). The results in table 3.5 are the

sum of times the test returned h=0 for the OPHL-I and OPHL-II, III case, for a total of 2184 runs (for each parameter).

Table 3.5: Intra-class comparisons for balanced speech, SK method, type I and II, III OPHL.

Feature	h=0	Feature	h=0	Feature	h=0
CPPs	291	STdB	83	MFCC7	42
HNR	174	MFCC3	76	MFCC6	41
SPI	118	SEn	71	MFCC12	38
MFCC1	116	MFCC13	66	MFCC9	37
RMS	112	MFCC10	61	f_0	33
MFCC2	86	MFCC11	43	MFCC5	31
SK	85	MFCC4	42	MFCC8	28

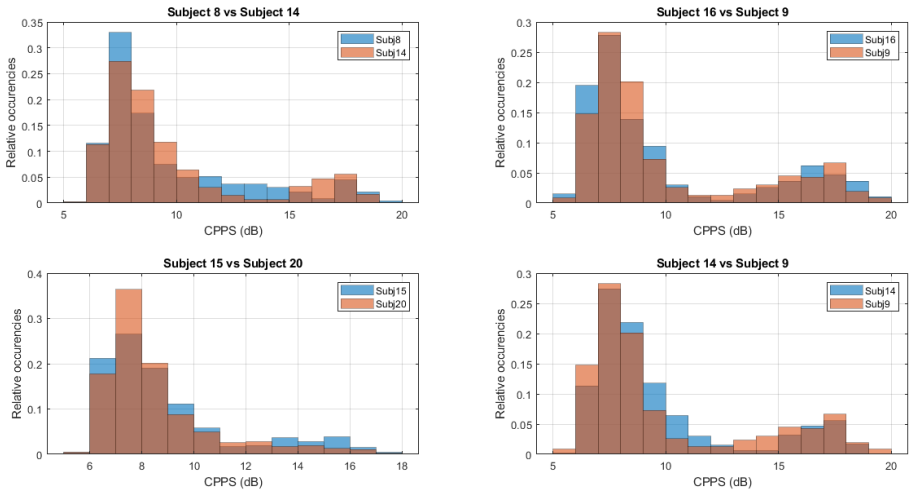


Figure 3.3: Distributions of CPPs values for OPHL-I, balanced speech, SK method

Recall that the total number of combinations in the intra-class comparison is obtained by calculating the binomial coefficient, comparing patients two by two at a time and there is a different number of patients for each data set. No characteristic features able to distinguish their own class were found, with a very feeble exception for the CPPs, which counted 291 out of 2184 occurrences. In figure 3.3 CPPs distributions among pairs of patients that underwent OPHL-I are reported

as an example, and they actually show common trends. Anyway, as the results in the following sections will show, CPPs almost never appeared in feature selection.

These results indicate strong variability among distributions of the same class, making it impossible to find a feature with a strong class recognition character. However, the results of the inter-class comparisons showed that the distributions have some features with good discriminatory power that allow the two classes to be distinguished.

3.2 Logistic Regression results

It is recalled that in the Feature Selection (FS) phase, the algorithm was trained with the combination of k features (k= 1, 2, 3, 4) and the FS algorithm indicated a combination of features for each k value; the following tables contain the cases thought to be most significant, indifferently with k= 1, 2, 3, 4 features.

3.2.1 Feature Selection results

During the Feature Selection (FS) phase, the algorithm automatically indicated a combination of features with the best accuracy obtained in the training phase and this was used to validate the algorithm. Checking carefully, however, there were cases where multiple feature sets provided the same maximum accuracy value. In this situation, validation was performed both with the combination of features returned by the algorithm with the highest accuracy and by selecting the combination of features with the highest AUC values among the ones with the highest accuracy.
In table 3.6 few examples cases of this scenario are reported. In detail, the top two couples refer to type I vs II, III OPHL, while the last couple refers to type II, III OPHL, 40 (I<5) vs 23 ($I \geq 5$).
The highest accuracy value in the validation phase of the model was obtained with the set of features with the highest AUC values; in particular, the best validation accuracy was reached in the balanced speech case, SK method, with an accuracy of 96.5%, by selecting the features SEn (95-th percentile), f_0(5,95-th percentile). In

Table 3.6: Examples where the FS algorithm identified multiple features with the same accuracy. The text in purple denotes the cases with the highest accuracy of the validated model.

Task	Selected Features	Method	AUC	Accuracy	Accuracy (valid)
bs	SEn (95-th percentile) f_0 (5-th percentile) f_0 (95-th percentile)	SK	0.99	98.8%	96.5%
bs	SEn (5-th percentile) f_0 (5-th percentile) f_0 (95-th percentile)	SK	0.96	98.8%	82.4%
sv	STdB (5-th percentile) HNR (median)	HNR	0.97	94.1%	92.9%
sv	MFCC5 (skewness) Jitta	HNR	0.96	94.1%	90.6%
bs	MFCC3 (skewness) CPPs (5-th percentile) f_0 (skewness)	SK	0.93	88.9%	84.1%
bs	f_0 (95-th percentile) f_0 (skewness) HNR (median)	SK	0.94	88.9%	84.1%

the last couple at the bottom, instead, the accuracy value after the model validation does not change according to AUC values.

It proves to be interesting the sustained vowel /a/ case, in which the feature selection algorithm went for Jitta, a perturbation parameter, and for MFCC5 (skewness); the choice of a statistics of MFCC5 confirmed the insight of the Kolmogorov-Smirnov test, that was pointed out in table 3.4.

Generally looking at the selected parameters, it should be noted the recurrence of parameter f_0 (5,95-th percentile, skewness) for the balanced speech cases. Moreover, the triplets of features selected in the first couple (balanced speech case, SK method) only differ for the SEn statistic, which switches from 5-th percentile to 95-th percentile.

By looking at the accuracy values relative to the training phase in table 3.6, it is possible to ascertain the LR model performed satisfactorily during the training phase of the model. The best results were obtained in type I vs II, III OPHL, in

particular for the balanced speech, SK method, with accuracy values of 98.8%. Nevertheless, an accuracy of 88.9% was obtained for the bottom couple.
In the balanced cases of classification based on index I (23, $I \geq 5$ vs 23, I<5 and 23, $I \geq 5$ vs 23, I<2.8) worst results were found than the previous case; as an example, the values in table 3.7.

Table 3.7: Accuracy values (pre, post validation) for the balanced data sets among patients who underwent type II, III OPHL (23, $I \geq 5$ vs 23, I<5 and 23, $I \geq 5$ vs 23, I<2.8).

Task	Selected Features	Method	Accuracy	Accuracy (valid)
sv	MFCC4 (skewness) SPI (range)	HNR	69.6%	67.4%
sv	MFCC10 (skewness) f_0 (95-th percentile)	HNR	80.6%	76.1%
bs	f_0 (5-th percentile) HNR (median)	SK	84.1%	82.6%
bs	f_0 (mean) f_0 (standard deviation) HNR (95-th percentile)	SK	93.5%	89.1%
bs	MFCC6 (skewness) CPPs (standard deviation) HNR (95-th percentile)	HNR	78.3%	73.9%

As far as concerns sustained vowel /a/ the accuracy values do not exceed 80.6% and in certain situations are below the 70%, as in the first top case. Regarding the balanced speech cases, the situation is slightly better, with accuracy values around the 90% and the best situation for SK method, trained with f_0 (mean, standard deviation), HNR (95-th percentile) which gave an accuracy of 93.5%. In all the reported cases, the results were lowered during the 5-fold cross-validation of the LR model; the highest accuracy of 89.1% is found in the aforementioned balanced speech case, with the combination of three features. Generally looking at the selected parameters for the balanced speech cases, the most common is f_0 (mean, 5,95-th percentile, median, standard deviation) and HNR (95-th percentile, median). The CPPs instead was selected only ones, with the standard deviation, a statistical measure of variability.

3.2.2 Best results of the Validation phase

The following tables summarize the best accuracy values obtained after the 5-fold cross-validation of the Logistic Regression model in the Classification Learner Tool, Matlab (R2022a). Specifically, only the validation outcomes for the models that obtained the highest accuracy values during the training phase of the LR model are reported, hence the cases for the balanced data sets (see table 3.7) among patients who underwent type II, III OPHL (23, $I \geq 5$ vs 23, I<5 and 23, $I \geq 5$ vs 23, I<2.8) were not taken into account.
In table 3.8 the OPHL-I vs II, III case is reported. This was the situation with the highest accuracy values, up to 96.5% for the SK method, selecting the features SEn (95-th percentile), f_0 (5-th,95-th percentile). This particular set of features was confirmed in the Kolmogorov-Smirnov inter-class tests results, as seen in section 3.1, table 3.1. Other statistics of SEn (median, standard deviation) were selected for the

Table 3.8: Best accuracy values of the model validation, case OPHL-I vs II,III.

Task	Selected Features	Method	AUC	TPR	TNR	Accuracy
bs	SEn (95-th percentile) f_0 (5-th percentile) f_0 (95-th percentile)	SK	0.96	96.8%	95.5%	96.5%
bs	f_0 (range) HNR (skewness)	HNR	0.96	96.8%	90.9%	94.1%
sv	SEn (median) SEn (standard deviation)	SK	0.90	96.8%	81.8%	92.9%
sv	STdB (95-th percentile) HNR (median)	HNR	0.97	96.8%	81.8%	92.9%
sv	MFCC1 (kurtosis) MFCC5 (skewness) HNR (5-th percentile) vAm	HNR	0.96	93.7%	81.8%	94.1%

sustained vowel /a/ case, SK method, obtaining an accuracy of 92.90%. The same accuracy value was obtained for another vowel /a/ case, HNR method, selecting STdB (95-th percentile) and HNR (median). In the case of the last row, vAM, a perturbation variability parameter, appears among the four selected features, and the validation accuracy is 94.1%. All Validation model shows AUC values above

0.90, indicating a good measure of separability and at least a 90% chance that the model will be able to distinguish between positive and negative class.
TPR values are 96.8% for all cases, except the last one, and are higher than TNR values; this highlights that these classifiers are more able in identifying patients with the worst vocal condition, and it is possible that a healthy patient would be incorrectly identified in the wrong class, creating a False Alarm.
In table 3.9, the values relative to OPHL-II,III with 23 ($I \geq 5$) and 40 (I<5, good quality of voice) are shown.

Table 3.9: Best classification metrics obtained with validation of the LR model, classification between OPHL-I vs II, III based on index I, case 40 (I<5) vs 23 ($I \geq 5$)case.

Task	Selected Features	Method	AUC	TPR	TNR	Accuracy
bs	f_0 (95-th percentile) f_0 (skewness) HNR (median)	SK	0.89	69.6%	92.5%	84.1%
bs	SEn (mean) StdB (5-th percentile)	HNR	0.77	60.9%	87.5%	77.8%
sv	MFCC3 (95-th percentile) MFCC5 (mode) MFCC7 (median)	SK	0.74	52.2%	87.5%	74.6%
sv	MFCC2 (mean) MFCC5 (mean) MFCC9 (range) MFCC13 (mean)	HNR	0.68	65.2%	75.0%	71.4%

At a glance, it can be seen that the accuracy values are ten percentage points lower than those of OPHL-I vs II, III cases, reaching 84.1% as the highest value. This was an expected result, since OPHL-I preserves almost the entire phonatory system, whereas OPHL-II, III are more invasive operations; it this therefore logical to assume that the classifier provides better performance in distinguishing classes with very different characteristics from a case where they have similar characteristics.
It is possible to notice how different statistics of Mel Coefficients MFCCs were always selected for the two sustained vowel /a/ cases reported.

As far as concerns the balanced speech task, SK method, as in the case in table 3.8, the f_0 (5, 95-th percentile) appears in the selected features, with HNR (median) obtaining an AUC value of 0.89, being the highest among these data set examples. The minimum AUC value is 0.68, in the case of the sustained vowel /a/, HNR method, with the mean of MFCC 2, 5, 13 and the range of MFCC 5, as selected features. MFCC5 and MFCC9 were confirmed in the execution Kolmogorov-Smirnov inter-class tests results, reported in section 3.1, table 3.3.

3.3 Uncertainty evaluation for the LR model

This section reports the most relevant results obtained with the method to obtain the Expanded Uncertainty $U(p)$ of the probability p returned by the validation of the LR model, described in Chapter 2.7.3. For convenience, the computation of the Expanded Uncertainty of the model was carried out on the LR models with k=2 features selected, having a total of three β coefficients. This trick made it possible to ease the calculation of sensitivity coefficients (via the partial derivative of the probability formula with respect to the β coefficients). The following sections report two cases that gave the best accuracy values during the validation phase, among the ones with k=2 selected features; they both refer to the type I vs II, III OPHL.

3.3.1 Balanced speech

The first model is a case of balanced speech, HNR method, with f_0 (range) and HNR (skewness) as selected features. Figure 3.4 shows the probabilities returned by the LR-validated model, as in the Classification Learner App, Matlab (R2022a), without their uncertainties. Blue elements are associated with Class 0, red ones are associated with Class 1. By looking at the graph, it should be noted some subjects whose probability values sway around the threshold at 0.5.
To evaluate the standard uncertainty $u(p)$ of the LR validated model, uncertainty propagation was implemented on the probability Eq. 2.22, in section 2.7.3:

$$\begin{aligned} p &= \frac{e^{\Theta^T \cdot x}}{1 + e^{(\Theta^T \cdot x)}} = \frac{1}{1 + e^{-(\Theta^T \cdot x)}} \\ \Theta^T x &= \beta_0 + \beta_1 x_1 + \beta_i x_2 + ... + \beta_N x_n \end{aligned} \tag{3.1}$$

All β coefficients are affected by uncertainty (SE) which are provided in the LR-validated model, returned by the function generated in the Classification Learner App, Matlab (R2022a).
Sensitivity coefficients were derived by unwinding the partial derivatives of the probability p with respect to the β coefficients. Subsequently, the standard uncertainty of the probability returned by the LR model was obtained from the uncertainty

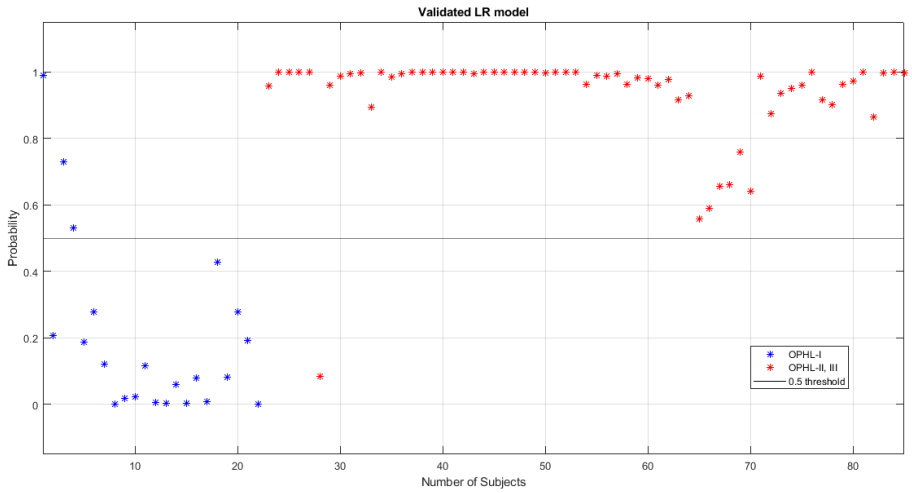


Figure 3.4: Probabilities returned by the LR-validated model without expanded uncertainty, balanced speech, HNR method, OPHL-I vs II, III.

propagation formula Eq. 2.29, in section 2.7.3:

$$u(p) = \sqrt{J_\beta \cdot COV_\beta \cdot J_\beta^T}$$
$$J_{i,j}(\beta) = \frac{\partial p_i}{\partial \beta_i}; j \in [1...N_F + 1]; i \in [1...N_S] \tag{3.2}$$

The propagation formula includes the sensitivity coefficients of the model and the variances and covariances of the β coefficients. The covariance values of the coefficients were extracted from the LR-validated model, returned by the function generated in the Classification Learner App, Matlab (R2022a). In the present case, the uncertainty formula in 3.2 was applied with the contributions of three sensitivity coefficients and three covariances. The expanded uncertainty $U(p)$ was obtained by multiplying the uncertainty $u(p)$ by a coverage factor 2; in this way, an interval of confidence for each probability value of the model was obtained, graphically represented by error bars in figure 3.5.

The uncertainty is directly related to sensitivity coefficients (see equation 2.28, section 2.7.3), which tend to increase for probability values around 0.5 and decrease for probabilities near 0 and 1; the graph in figure 3.5 confirmed the insight. In the figure it should be noted some elements whose error bars intersect the threshold

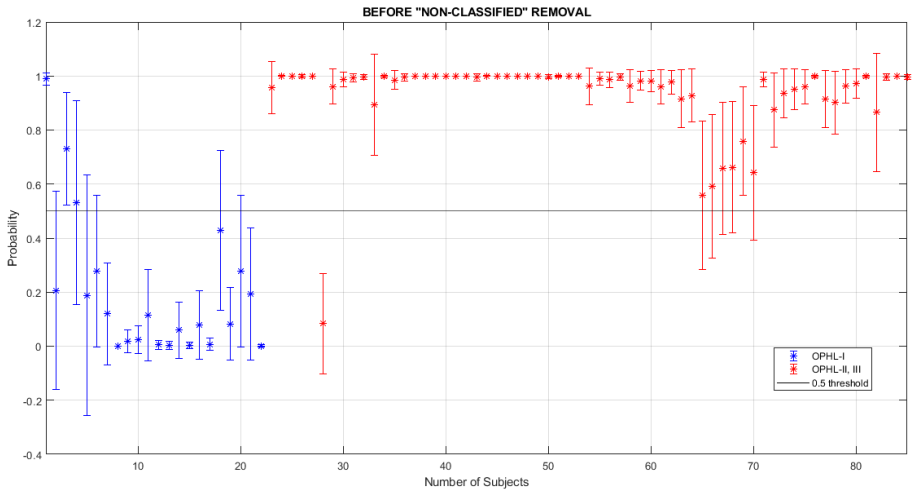


Figure 3.5: Probabilities returned by the LR-validated model with expanded uncertainty, balanced speech, HNR method, OPHL-I vs II, III.

at 0.5; in particular, six subjects for Class 0 and five subjects for Class 1. The LR-validated model classified all subjects regardless of their uncertainty values; all of them were classified as TP and TN, beside subject 4, who was classified as FP (see confusion matrix in fig. 3.6).

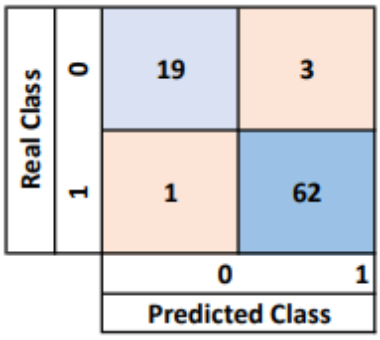


Figure 3.6: CM, balanced speech, HNR method, OPHL-I vs II, III.

The expanded uncertainty associated with these subjects suggested that their probability values were too doubtful to be considered in the binary classification. Hence, the third class of "non-classified" was introduced. For the case in point, the "non-classified" subjects were a total of 11, so the total number of subjects held to be classifiable changed from 85 to 74, as shown in fig. 3.7.

To have an objective perception of the effect of "non-classified" subjects on the

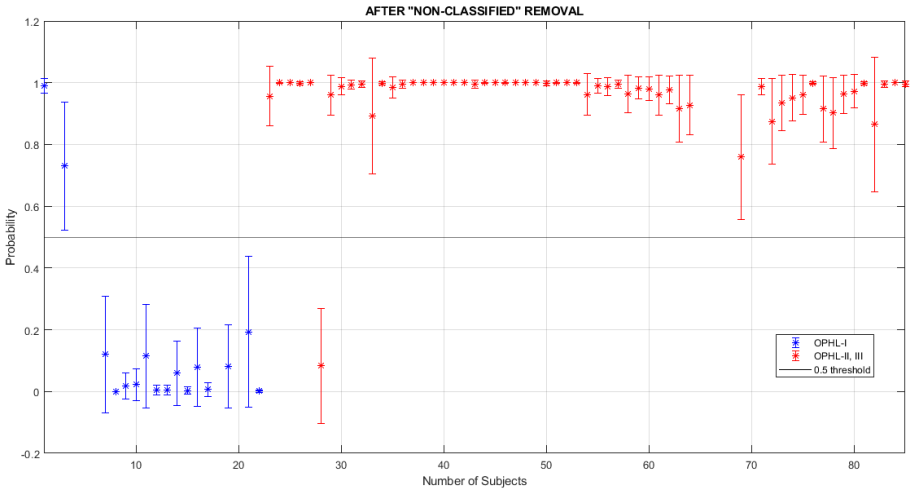


Figure 3.7: Probabilities returned by the LR-validated model with expanded uncertainty, balanced speech, HNR method, OPHL-I vs II, III, after the removal of "non-classified".

overall performance of the classifier, new metrics were introduced, such as Realistic Accuracy (Acc_{real}) and Fraction of Classified (FoC).

The Realistic Accuracy is nothing but the calculation of the accuracy (as reported in Eq. 2.23) by excluding the elements belonging to the "non-classified" class; hence considering the elements in the confusion matrix in fig. 3.8.

The Realistic Accuracy value resulted to be 95.9%, improving of about two percentage points the accuracy value of 94.1% provided by the original LR model.

The fraction of Classified instead is an index of the number of elements labeled as "non-classified" and resulted to be 0.87. In table 3.10 the results of the Uncertainty evaluation are summarized.

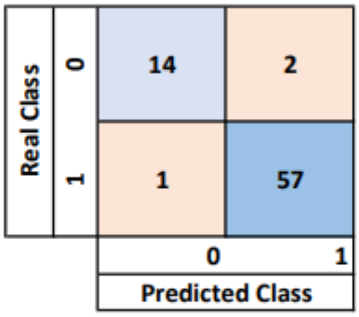


Figure 3.8: CM, balanced speech, HNR method, OPHL-I vs II, III, after the removal of "non-classified".

Table 3.10: Summary of the evaluation metrics before and after the removal of "non-classified" subjects.

TPR	TNR	Acc	TPR	TNR	Acc_{real}	FoC
96.8%	90.9%	94.1%	98.3%	87.5%	95.9%	0.87

3.3.2 Sustained vowel /a/

The second model refers to a case of uttering of the sustained vowel /a/, SK method, with SEn (median, standard deviation) as selected features. Similar considerations to the ones reported in the previous section were made.
In fig. 3.9 the probabilities returned by the LR-validated model are shown. It

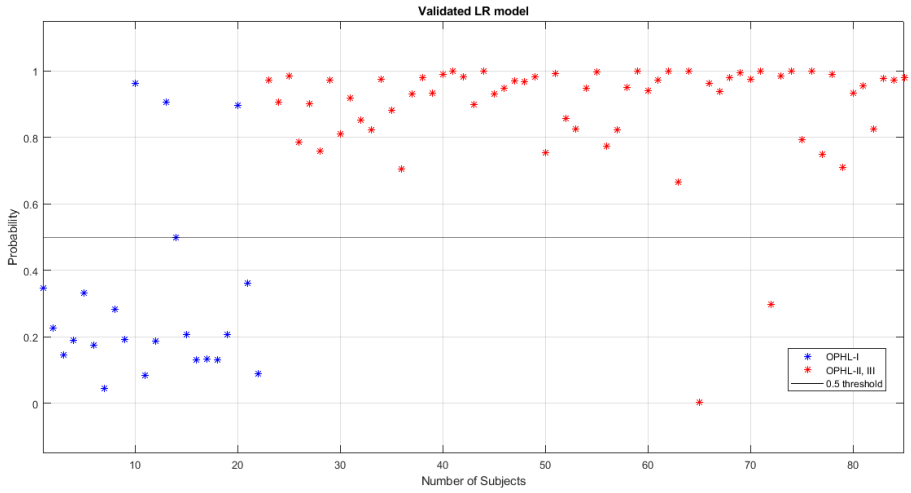


Figure 3.9: Probabilities returned by the LR-validated model without expanded uncertainty, sustained vowel /a/, SK method, OPHL-I vs II, III.

already jumps to the eye the presence of a doubtful subject, whose probability value is on the threshold of 0.5; this subject is part of Class 0 but was assigned by the LR model to Class 1, creating a FP (see confusion matrix in fig. 3.10).
Figure 3.11 reports the probabilities with the computed expanded uncertainties; the doubtful subjects with probabilities values around 0.5 were five for Class 0 and three for Class 1. As expected, the aforementioned subject on the threshold at 0.5 has the widest uncertainty among the other subjects.
After the removal of the "non classified" subjects, shown in figure 3.12 the total number of subjects was reduced from 85 to 77. The elimination of the critical patient on the threshold value, which carried a strong uncertainty component, remodeled the number of FP by one (see confusion matrix in fig. 3.13).

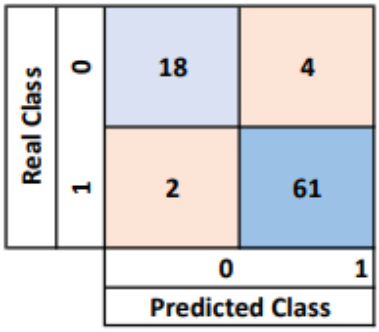


Figure 3.10: CM, sustained vowel /a/, SK method, OPHL-I vs II, III.

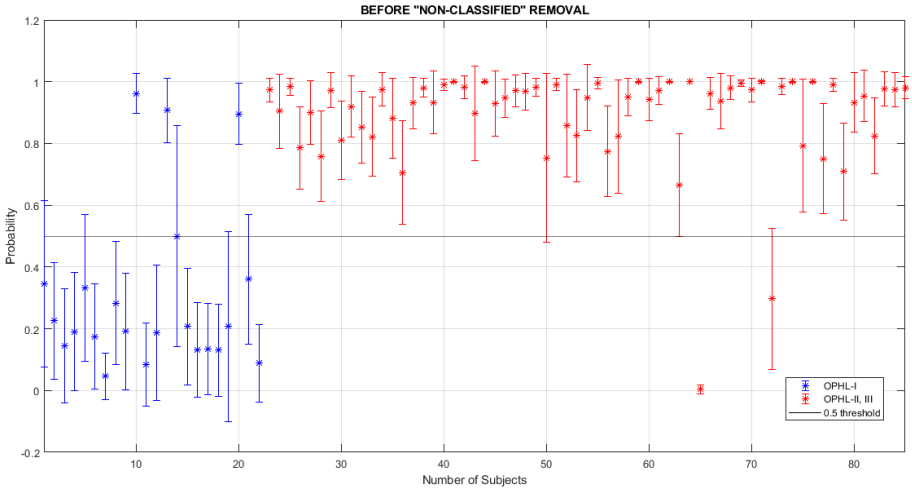


Figure 3.11: Probabilities returned by the LR-validated model with expanded uncertainty, sustained vowel /a/, SK method, OPHL-I vs II, III.

The model provided a Realistic Accuracy of 94.8%, increasing with respect to the accuracy value of 92.9% provided by the original LR model. The Fraction of Classified turned out to be 0.91. In table 3.11 the results of the Uncertainty evaluation are summarized.

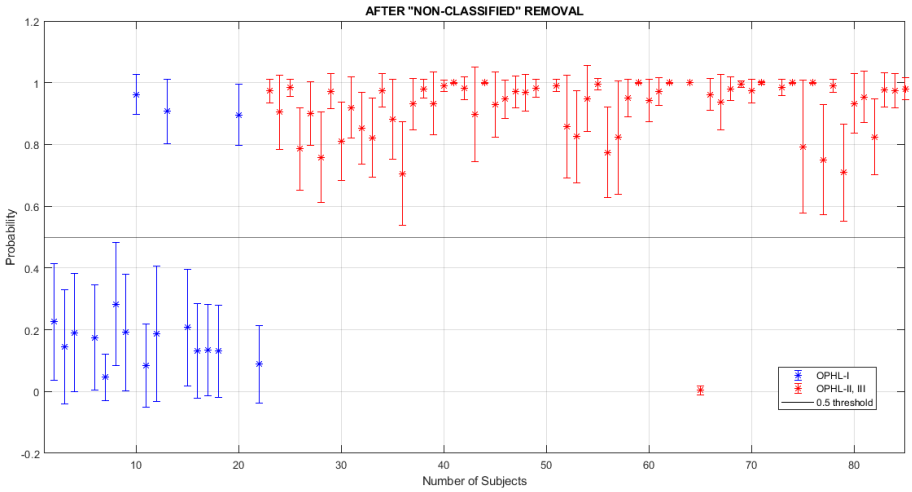


Figure 3.12: Probabilities returned by the LR-validated model with expanded uncertainty, sustained vowel /a/, SK method, OPHL-I vs II, III, after the removal of "non-classified" subjects.

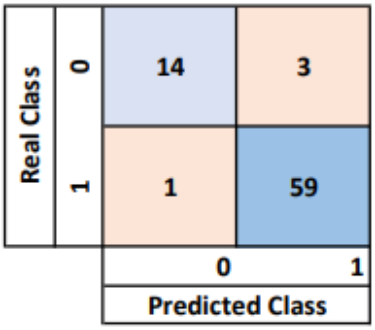


Figure 3.13: CM, sustained vowel /a/, SK method, OPHL-I vs II, III, after the removal of "non-classified" subjects.

Table 3.11: Summary of the evaluation metrics before and after the removal of "non-classified" subjects.

TPR	TNR	Acc	TPR	TNR	Acc_{real}	FoC
96.8%	81.8%	92.9%	98.3%	82.4%	94.8%	0.91

Chapter 4

Conclusions

In this work, two different vocal tasks (reading of a phonetically balanced text and vocalization of the sustained vowel /a/) of 85 subjects who underwent open partial horizontal laryngectomy were examined, to identify vocal features that are representative of the vocal quality of substitution voices. These features should allow clinicians to have an objective assessment of both post-surgery phonatory impairment and the effectiveness of rehabilitation.

The available vocal material was pre-processed according to two different methods, which relied on the Harmonic-to-Noise Ratio (HNR) and the Spectral Kurtosis (SK), in order to select the harmonic frames that were used to extract vocal features in the time, spectral and cepstral domains. The subjects were classified according to two criteria: the first based on the severity of the surgery (22 OPHL-I subjects vs 63 OPHL-II, III subjects), the second based on the voice quality according to index I (Intelligibility) of the INFVo scale, among patients with the worst vocal conditions (23 OPHL-II, III subjects with index I ≥ 5 vs 40 OPHL-II, III subjects with index I < 5). The discriminatory power of the extracted features was initially assessed using the two-sample Kolmogorov-Smirnov test, which made it possible to compare the distributions of the extracted features between each pair of subjects. The outcomes of the test indicated the features Harmonic-to-Noise Ratio (HNR), fundamental frequency (f_0), Mel Frequency Cepstral Coefficients ($MFCCs$), Spectral Entropy (SEn), and Spectral Kurtosis (SK) as the best candidates to assess the quality of substitution voices. Some of the results obtained in the previous phase were

confirmed with an analysis based on the classification performance of a Logistic Regression (LR) model, trained using a single feature or a combination of 2, 3, and 4 uncorrelated features. Out of the trained models, the combination of features (or the single feature) that provided the best classification performance (in terms of accuracy) for each of the tested comparisons was selected. The selected features were used to validate the LR model with 5-fold cross-validation, directly in the Classification Learner App, Matlab (R2022a). If there were multiple features with the same maximum accuracy value, validation was performed with both the features selected by the algorithm and the features with the largest area under the curve (AUC), among those with the highest accuracy. However, the results showed that this detail turned out to be rather irrelevant to the performance of the validated model.

The best classification results achieved, summarized in the previous chapter, make it clear that the balanced speech task is the most suitable vocal material for assessing the quality of substitution voices. In addition, the SK criterion was proven to be the most effective method for selecting harmonic frames from patients' voice recordings. In the comparison between $\mathrm{OPHL}-\mathrm{I}$ and $\mathrm{OPHL}-\mathrm{II,III}$ subjects (balanced speech, SK method), a classification accuracy of 96.5% (sensitivity 96.8% and specificity 95.5%) was obtained using a 5-fold cross-validation technique to prevent over-fitting phenomena. The selected features were f_0 (5-th percentile),f_0 (95-th percentile) and SEn (95-th percentile). As expected, worse results were obtained in the comparison between the classes $\mathrm{OPHL}-\mathrm{II,III(I}<5)$ and $\mathrm{OPHL}-\mathrm{II,III(I}\geq 5)$, as both categories include patients who underwent a very severe surgery. Again, the balanced speech task, pre-processed through the SK method, provided the highest accuracy of 84.1% (sensitivity 69.6% and specificity 92.5%) using a 5-fold cross-validation technique. The selected features were f_0 (95-th percentile), f_0 (skewness) and HNR (median).

Eventually, to evaluate the classification performance more realistically, a procedure that relies on the confidence interval evaluation of the probability provided by the logistic regression model was proposed. As a first thing, the sensitivity coefficients were derived as the partial derivatives of the probability function with respect to the model coefficients. The interval was then created by applying a coverage factor 2 to the standard uncertainty, previously estimated with the uncertainty propagation

formula, with the contribution of the variances and covariances of the model coefficients. Confidence intervals with values that intersected the threshold set at 0.5 suggested that their probability values were too doubtful to be considered in the binary classification, hence the third class of "non-classified" was introduced. To get objective feedback on the effect of "non-classified" subjects on overall classification performance, the new metrics Realistic Accuracy (Acc_{real}) and Fraction of Classified (FoC) were created. By way of example, the balanced speech task, pre-processed through the HNR method, with f_0 (range) and HNR (skewness) as selected features, provided a Fraction of Classified of 0.87 and a Realistic Accuracy of 95.9%, improving of about two percentage points the original accuracy value of 94.1%. This procedure can clearly be extended to any type of study involving classification based on a logistic regression model.

To clarify, this thesis work was done with a view to providing an objective support to clinicians in evaluating the rehabilitation pathway of post-laryngectomized patients and is in no way intended to be a substitute for conventional methods. Future developments could certainly involve expanding the dataset analyzed so that the characteristics identified as representative can be verified with greater certainty.

Bibliography

[1] *How does my voice works?* URL: https://www.templehealth.org/about/blog/how-does-my-voice-work (cit. on p. 2).

[2] *Voice Biometrics Technologies and Applications for Healthcare: an overview - Scientific Figure on ResearchGate*. accessed 12 Nov, 2022. URL: https://www.researchgate.net/figure/The-peripheral-phonation-system-Purves-2012_fig2_329059822 (cit. on p. 3).

[3] *Laryngeal Cartilages*. URL: https://teachmeanatomy.info/neck/viscera/larynx/laryngealcartilages/ (cit. on pp. 3, 4).

[4] Zhaoyan Zhang. «Mechanics of human voice production and control». In: *The journal of the acoustical society of america* 140.4 (2016), pp. 2614–2635 (cit. on pp. 4, 9).

[5] Kenneth Beare. «"Voiced vs. Voiceless Consonants"». In: (august 29 2020). URL: thoughtco.com/voiced-and-voiceless-consonants-1212092 (cit. on p. 4).

[6] Wikimedia Commons contributors. Date of last revision: 20 April 2021 04:53 UTC. URL: https://commons.wikimedia.org/w/index.php?title=File:Gray956.png&oldid=554063012 (cit. on p. 5).

[7] Juan Suarez-Quintanilla, Alejandro Fernandez Cabrera, and Sandeep Sharma. «Anatomy, head and neck, larynx». In: *StatPearls [Internet]*. StatPearls Publishing, 2021 (cit. on p. 5).

[8] *Tumori della laringe*. URL: https://www.humanitas.it/malattie/tumori-della-laringe/ (cit. on p. 5).

[9] Giovanni Succo, Giuseppe Rizzotto, Erika Crosetti, Marco Lucioni, Patrizia Olmi, and Lisa Licitra. «Il carcinoma laringeo». In: *Tumori della testa e del collo*. Springer, 2011, pp. 331–360 (cit. on p. 6).

[10] Antonio Schindler, Nicole Pizzorni, Marco Fantini, Erika Crosetti, Andy Bertolin, Giuseppe Rizzotto, and Giovanni Succo. «Long-term functional results after open partial horizontal laryngectomy type IIa and type IIIa: A comparison study». In: *Head & Neck* 38.S1 (2016), E1427–E1435 (cit. on p. 6).

[11] V Di Nicola, ML Fiorella, DA Spinelli, and R Fiorella. «Acoustic analysis of voice in patients treated by reconstructive subtotal laryngectomy. Evaluation and critical review». In: *Acta otorhinolaryngologica italica* 26.2 (2006), p. 59 (cit. on p. 7).

[12] Youri Maryn and Nelson Roy. «Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity». In: *Jornal da Sociedade Brasileira de Fonoaudiologia* 24 (2012), pp. 107–112 (cit. on p. 7).

[13] AL Webb, PN Carding, IJ Deary, K MacKenzie, N Steen, and JA Wilson. «The reliability of three auditory perceptual scales for dysphonia». In: *European Archives of Oto-Rhino-Laryngology* (2004) (cit. on p. 7).

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

[15] Mieke Moerman, Jean-Pierre Martens, Lise Crevier-Buchman, Else de Haan, Stephanie Grand, Christophe Tessier, Virginie Woisard, and Philippe Dejonckere. «The INFVo perceptual rating scale for substitution voicing: development and reliability». In: *European Archives of Oto-Rhino-Laryngology and Head & Neck* 263.5 (2006), pp. 435–439 (cit. on pp. 7, 14, 15).

[16] A Schindler, D Ginocchio, M Atac, P Maruzzi, S Madaschi, F Ottaviani, and F Mozzanica. «Reliability of the Italian INFVo scale and correlations with objective measures and VHI scores». In: *Acta otorhinolaryngologica italica* 33.2 (2013), p. 121 (cit. on pp. 7, 8).

[17] Renée Speyer, Hans CA Bogaardt, Valéria Lima Passos, Nel PHD Roodenburg, Anne Zumach, Mariëlle AM Heijnen, Laura WJ Baijens, Stijn JHM Fleskens, and Jan W Brunings. «Maximum phonation time: variability and reliability». In: *Journal of Voice* 24.3 (2010), pp. 281–284 (cit. on p. 8).

[18] R Kazi, J De Cordova, A Singh, R Venkitaraman, CM Nutting, P Clarke, P Rhys-Evans, and KJ Harrington. «Voice-related quality of life in laryngectomees: assessment using the VHI and V-RQOL symptom scales». In: *Journal of voice* 21.6 (2007), pp. 728–734 (cit. on p. 8).

[19] Eva Villanueva, Maria Paula Fernández, Giovanna Arena, José L Llorente, Juan P Rodrigo, Fernando López, and César Álvarez-Marcos. «Validation of "Self-Evaluation of Communication Experiences after Laryngectomy"(SECEL) Questionnaire for Spanish-Speaking Laryngectomized Patients». In: *Cancers* 14.14 (2022), p. 3347 (cit. on p. 8).

[20] Jonathan Delgado-Hernández, Nieves M León-Gómez, Laura M Izquierdo-Arteaga, and Yanira Llanos-Fumero. «Cepstral analysis of normal and pathological voice in Spanish adults. Smoothed cepstral peak prominence in sustained vowels versus connected speech». In: *Acta Otorrinolaringologica (English Edition)* 69.3 (2018), pp. 134–140 (cit. on pp. 9, 10).

[21] Alessio Carullo, Alessio Atzori, Lorenzo Midolo, Alberto Vallan, Marco Fantini, and Giovanni Succo. «Rehabilitation Monitoring of Post-Laryngectomy Patients through the Extraction of Vocal Parameters». In: *2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. 2022, pp. 1–6. DOI: 10.1109/MeMeA54994.2022.9856487 (cit. on pp. 9, 12, 20, 29).

[22] 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. 11).

[23] Rubén Fraile and Juan Ignacio Godino-Llorente. «Cepstral peak prominence: A comprehensive analysis». In: *Biomedical Signal Processing and Control* 14 (2014), pp. 42–54 (cit. on pp. 11, 12).

[24] Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch, and Sara Sandabad. «Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease». In: *2015 International conference on electrical and information technologies (ICEIT)*. IEEE. 2015, pp. 300–304 (cit. on p. 12).

[25] *Phonetic balance*. URL: https://medical-dictionary.thefreedictionary.com/phonetic+balance (cit. on p. 14).

[26] *Software Instruction Manual "Multi-Dimensional Voice Program (MDVP)" Model 5105*. KayPENTAX. Chap. Appendix C, pp. 135–189 (cit. on p. 17).

[27] Zbynık Tychtl and Josef Psutka. «Speech production based on the mel-frequency cepstral coefficients». In: *Sixth European Conference on Speech Communication and Technology*. 1999 (cit. on pp. 18, 19).

[28] Achraf Benba, Abdelilah Jilbab, Ahmed Hammouch, and Sara Sandabad. «Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease». In: *2015 International conference on electrical and information technologies (ICEIT)*. IEEE. 2015, pp. 300–304 (cit. on p. 19).

[29] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, M. Spadola Bisetti, and J. Colombini. «Cepstral peak prominence smoothed distribution as discriminator of vocal health in sustained vowel». In: *2017 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. 2017, pp. 1–6. DOI: 10.1109/I2MTC.2017.7969748 (cit. on p. 25).

[30] João Paulo Teixeira, Carla Oliveira, and Carla Lopes. «Vocal acoustic analysis–jitter, shimmer and hnr parameters». In: *Procedia Technology* 9 (2013), pp. 1112–1122 (cit. on pp. 25, 27, 28).

[31] Alessio Carullo, Adriano Anibaldi, Arianna Astolfi, Alessio Atzori, Viviana Cennamo, and Giovanni Zito. *A New Paradigm of Effective Communication based on Voice Shapes*. Universitätsbibliothek der RWTH Aachen, 2019 (cit. on p. 30).

[32] Yingyong Qi and Robert E Hillman. «Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals». In: *The Journal of the Acoustical Society of America* 102.1 (1997), pp. 537–543 (cit. on p. 31).

[33] Frank J Massey Jr. «The Kolmogorov-Smirnov test for goodness of fit». In: *Journal of the American statistical Association* 46.253 (1951), pp. 68–78 (cit. on p. 32).

[34] Alessio Atzori. «Effects of measurements uncertainty on classification algorithms, as applied to vocal features for health assessment and early diagnosis». PhD thesis. ScuDo, 2022 (cit. on pp. 34, 40).