POLITECNICO DI TORINO

Master Degree in Biomedical Engineering

Vocal-load and vocal-health assessment based on the estimation of acoustic parameters of the vocal signal

Supervisor

Prof. Alessio Carullo

Candidate

Chiara Gervasi

Co-Supervisors

Prof. Alberto Vallan

Eng. Antonella Castellana

Academic Year 2017-2018

Abstract

Humans rely on their voice to inform, persuade and interact with other people. For this reason, a phonation disorder can be very limiting, interfering with the ability to communicate and with normal daily activities; it has a negative impact on human life from a physical, social, emotional and economic point of view. People develop voice disorders for a variety of causes, which can be related to the improper use of vocal folds, affecting the vocal load, or to dysfunctions of the voice apparatus, undermining the vocal health. For this reason, the need to find new approaches and methods for evaluating the voice status, both qualitative and quantitative, has acquired an increasing resonance over the time. In detail, in the last few years, thanks to in-field voice monitoring, the objective analysis of laryngeal problems has been spreading: it consists in the estimation of acoustic parameters from the vocal signal, which are able to quantify vocal load and vocal health. Therefore, the overall aim of this study is the search for systems and techniques based on the voice acoustic analysis as tools to prevent vocal disorders and assess their severity.

The first part of this thesis focuses on the classification between healthy and pathological voices. Voice samples have been gathered thanks to the collaboration with the phoniatrics department at San Giovanni Battista Hospital, in Turin: voluntary patients and control subjects undertook the experiments, after the videolaryngostrobscopy examination and the voice perceptual evaluation performed by the physician. For recordings, different speech materials and devices were used. The protocol consisted of three sustained vowel /a/, the reading of a phonetically balanced passage and a free speech, and the subject was equipped with a microphone in air and two contact microphones. At the end of the medical examination, subjects filled in a voice self-assessment questionnaire, called “Profilo di Attività e Partecipazione Vocale” (PAPV). The signal processing started from the awareness of promising results obtained by previous studies, based on the Cepstral Peak Prominence Smoothed distribution and its descriptive statistics as indicators of vocal condition. As a result, at first, existing single-variable logistic regression models, for the different devices, have been

validated applying them to a new data set, in order to evaluate the generalization ability of the classifiers. Then, only for sustained vowels, perturbation parameters and HNR were implemented and included in the statistical analysis; by contrast, in reading and free speech only CPPS statistics have been estimated. Differently from earlier experiments, for all the speech materials two-variable logistic regression models were tested, in order to combine information from different parameters. The performances of the tested models have been evaluated and compared to each other, mainly through their Receiver Operating Characteristic (ROC) and the relative area (AUC). As regards sustained vowel, results identified as the best model the one composed by CPPS 5^{th} percentile and PPQ perturbation parameter for microphone in air and piezoelectric microphone, exhibiting an AUC of 0.92 and 0.90 respectively, and the one composed by CPPS standard deviation and PPQ for Electret Condenser Microphone, with an AUC equal to 0.85. On the other hand, in reading task were found the following models: CPPS range - CPPS mean (AUC equal to 0.88) and CPPS 5^{th} percentile – CPPS median (AUC equal to 0.80) for microphone in air and ECM, respectively. For each found model, a threshold has been selected to accomplish the classifier building; for this purpose, slight priority has been given to sensitivity.

People who use their voice professionally often are subjected to voice disorders that are usually caused by vocal fold hyperfunction, which consists in vocal abuse leading to vocal load increase. According to several statistics, one of the largest categories of professional voice users are teachers. In presence of adverse environmental conditions, such as background noise, teachers are inclined to increase their voice and speak with higher vocal loudness resulting in vocal effort. Therefore, the second part of this thesis deals with experiments on vocal load parameters and their changes with background noise levels in primary school teachers. Long-term monitorings during teaching hours have been performed involving seven teachers of the primary school Roberto D'Azeglio, in Turin, using a piezoelectric microphone. The vocal activity was monitored during two different time periods, at a distance of about one month and half from each other. Furthermore, another distinction has been done between recordings before and

after recreation time. Only plenary lesson portions of the monitorings have been considered, in order to reduce variability for the next processing. At the same time, background noise levels have been detected during the lessons and the LA90 distributions have been extracted. After a preliminary calibration procedure, for each acquisition, the parameters that describe vocal load have been calculated: mean, median and standard deviation of Sound Pressure Level (SPL) and F0 distributions and the voicing time percentage (Dt%). Then, the relationship between these parameters and noise levels have been investigated in the several groups of voice samples. Regarding the differences between the two time periods, results did not lead to a particular proof, differently from the differences revealed between before and after recreation time, where the variations of the parameters indicate an increased vocal load: LA90 median rose by 3 dB, F0 mean by 12 Hz and SPL mean by 1 dB, on average. Finally, vocal effort has been evaluated through the parameter SPL equivalent (at 1 m from the speaker's mouth), proving a significant increase for both the comparison: vocal effort of "shout" type ($SPL_{eq,1m} \geq 78dB$) appeared. Such outcomes are promising and support those studies that aim to find a system able to quantify vocal fatigue and thus identify the risk of vocal dysfunction in professional voice users.

Contents

1. Introduction

1.1 Vocal apparatus and phonation

The human vocal apparatus includes the lungs as a source of air, the vocal folds in the larynx put in vibration and a series of resonant chambers, which are the pharynx, the mouth and the nasal cavities (fig. 1.1). All these components work together leading to phonation, i.e. voice production. The lungs can be described as the “generator”, since they provide the necessary airflow: during speaking, the air expelled from the lungs moves up through the trachea to the larynx. The latter consists of a set of muscles and pieces of cartilage, with variable degrees of mobility, which can be raised or lowered like a gate to protect bronchi and lungs from food and other foreign bodies. Then, in the larynx, the air passes over the vocal folds. These folds are a matched pair of muscles and ligaments, pearly white in colour and coated with mucus; they are attached horizontally from the thyroid cartilage at the front to the arytenoid cartilages at the rear. During breathing, the vocal cords are completely separated and relaxed; by contrast, in speech the larynx cartilages press them against each other, thus closing the opening between them, known as the glottis. Under the pressure of the air being exhaled, the vocal folds separate and close again immediately, causing the air pressure beneath the glottis to increase again.

By opening and closing the glottis rapidly during phonation, the vocal folds release the air from the lungs in a vibrating flow, so that the acoustic vibrations, the sounds that are the raw materials for the words themselves, are produced. This first section between the lungs and the glottis is defined as “glottis tract” and represents the source in the phonation. It is necessary to transform these sounds into words: they are shaped by the rest of the vocal apparatus, the “vocal tract”, which starts after the vocal cords up to the lips, including the nasal cavities. This section acts as a “resonator”, a complex filter that alters the sounds issuing from the glottis, amplifying some frequencies while attenuating others. Therefore,

while the larynx produces the vibrations without which the voice would not exist, it is these other parts of the phonetic apparatus that make the voice so flexible and versatile. The soft palate, tongue, teeth, lips, and others parts of the mouth modulate the sound varying their position, so that vowels and consonants can be produced.

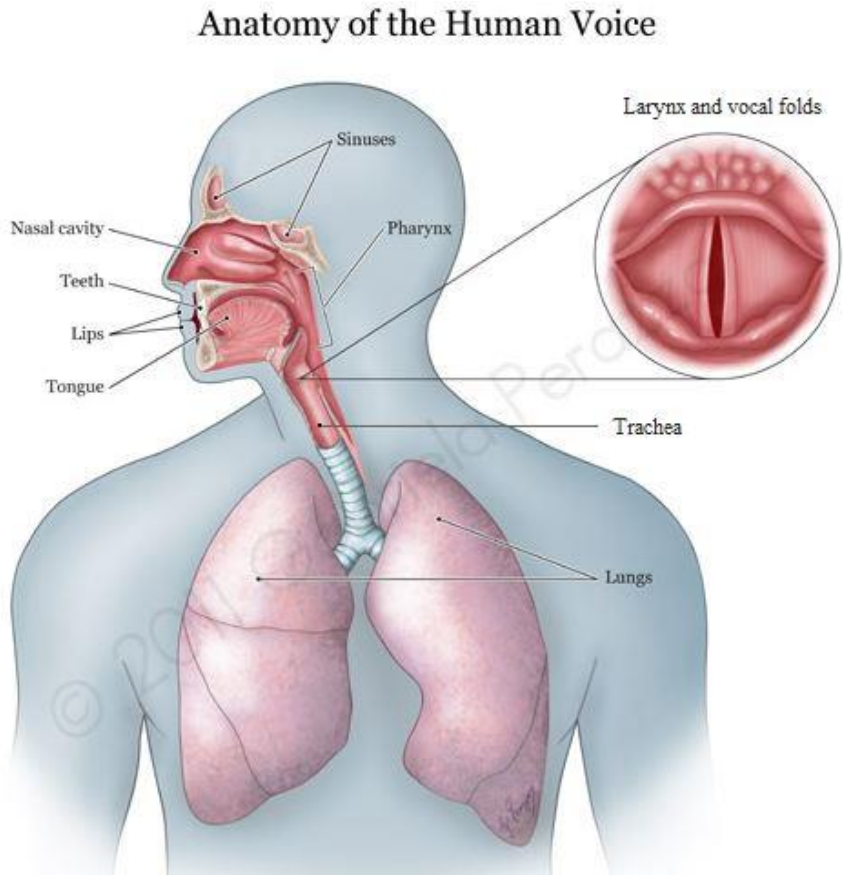


Figure 1.1: Diagram of the human vocal apparatus.

Certainly, the vocal folds play a central role in the voice production (fig. 1.2.). These elastic elements constitute the "vibrator" component of the phonetic apparatus. From the physiological point of view, the vibration frequency is directly proportional to the elastic characteristics of the vocal cords, which vary with the state of tension, and inversely proportional to the mass and length of the same. In particular, the length of the vocal folds in an adult man is 17-25 mm, in an adult woman is 12-17 mm. The structure, morphology and mechanical

properties of the vocal folds regulate the voice quality: variations in their mass and in their length can occur, due to structural alterations of the vocal folds, such as the presence of edema, nodules, polypes, etc.

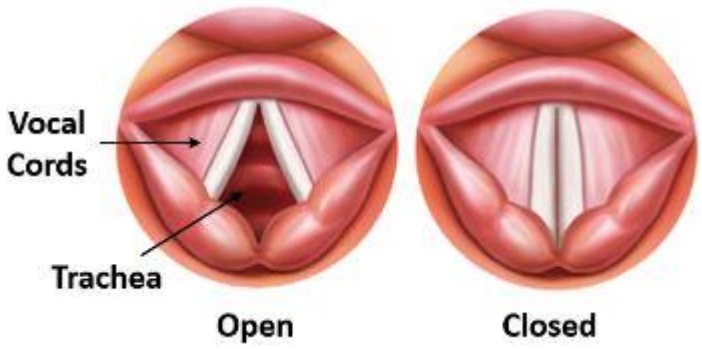


Figure 1.2: Vocal folds.

1.2 Acoustic characteristics of the vocal signal

As described in the previous paragraph, the phonation is the production of an acoustic signal from a source, which essentially consists of the vocal folds. Under normal conditions and from a physical point of view, the glottal signal, not yet filtered by the vocal tract, is a quasi-periodic complex signal [1]. As a complex signal, it is the algebraic sum of a series of sinusoidal signals, called spectral components; each component is characterized by its own frequency, intensity and phase. If the signal is periodic complex with fundamental period T_0, the components, defined also harmonics, have frequencies that are integer multiples of the fundamental frequency $F_0 = 1/T_0$. The latter, for the laryngeal signal coincides with the frequency of opening and closing of the glottis, i.e. the frequency of vibration of the vocal cords. The term "quasi" suggests that the characteristics of frequency and amplitude of the signal can change over time. Short-term perturbations could take place, observable from a period to the following one; for the fundamental frequency, these variations can be of the order of +/- 25 Hz and are necessary to provide naturalness to the speech [1]. There can

be long-term perturbations, differences between the start and the extinction of the sound emission, or to achieve special intonations, or to convey interpersonal attitudes. As it will be explained in the next paragraph, these variations in amplitude and fundamental frequency can characterize a pathological voice, if they are pushed beyond a certain normative value. Overall, observing few adjacent periods of the laryngeal signal, the complete contact of the vocal folds during the phonation ensures the production of an acoustic signal with high periodicity. Moreover, the value of F_0 is affected by several factors, such as length, tension level and mass of vocal cords; for example, it increases as tension and stiffness increase and decreases as vocal fold dimensions increase. In detail, shorts cords have high vibration frequencies. The F_0 variations fluctuate around a mean value, which is distinctive for each individual. This varies according to age, gender and type of vocal activity. The F_0 average ranges from 255 to 440 Hz for children, from 175 to 245 Hz for female adults and from 105 to 160 Hz for male adults [1].

Once come out of the vocal cords, the glottis signal is subjected to the filtering action of the vocal tract, which turns it into the vocal signal. However, F_0, the first harmonic of the glottal signal, remains the same also in the vocal signal; only the amplitude of the spectral components is modified. Figure 1.3 allows to compare the characteristics of the glottis signal, acquired with a contact microphone, and those of the corresponding vocal signal, emitted at the level of the lips and acquired with a microphone in air.

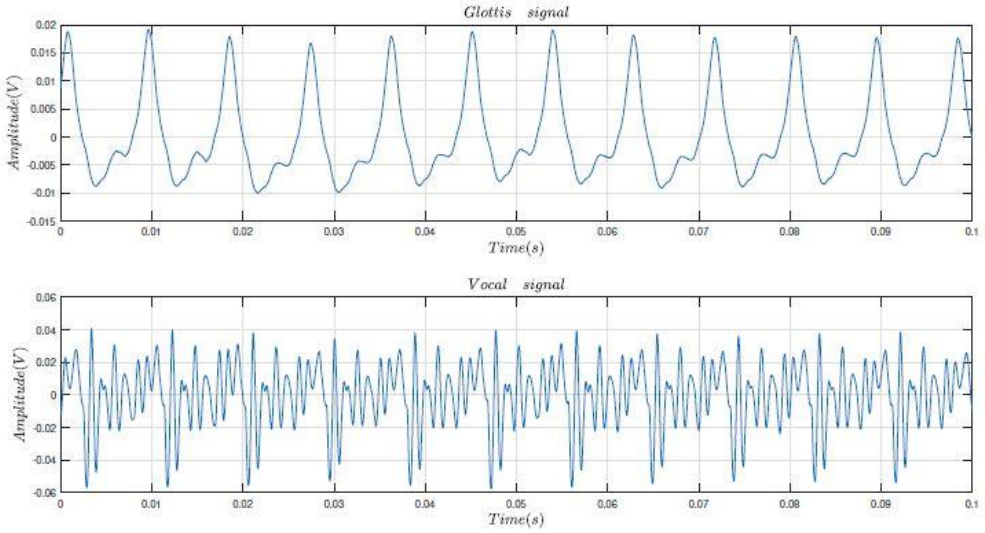


Figure 1.3: Glottis signal compared to vocal signal in a sustained vowel /a/.

Both signals are related to a sustained vowel /a/. In particular, it is possible to notice that:

- The glottis signal has a simpler waveform than the vocal signal.
- The glottis signal keeps approximately equal, unlike the vocal signal that vary considerably from a period to the other.

Figure 1.4 shows the power spectrum of the laryngeal and vocal signals highlighting the filter effect of the vocal tract, which is made up of laryngeal cavity, pharynx, oral cavity and nasal cavity. This structure forms the articulatory system: it exhibits variable volume and shape, so that the different ways of moving of the components, i.e. articulation, allow the acoustic signal to comprise much more information (vowels). In particular, the peaks of the spectral envelope of the filtered signal, which correspond to the peaks present in the filter transfer function, are called Formants, indicated sequentially with F_1, F_2,… They are the harmonics of the vocal signal with the maximum energy and represent the resonant frequencies of the vocal tract. Finally, the vocal energy covers a frequency band that exceeds 10 KHz. However, the band of interest in the analysis of the vocal signal reaches about 5 KHz, as a result of the recording equipment

[1]. Consequently, for an analytical evaluation of the recorded signal, it is necessary to know the frequency response of the device.

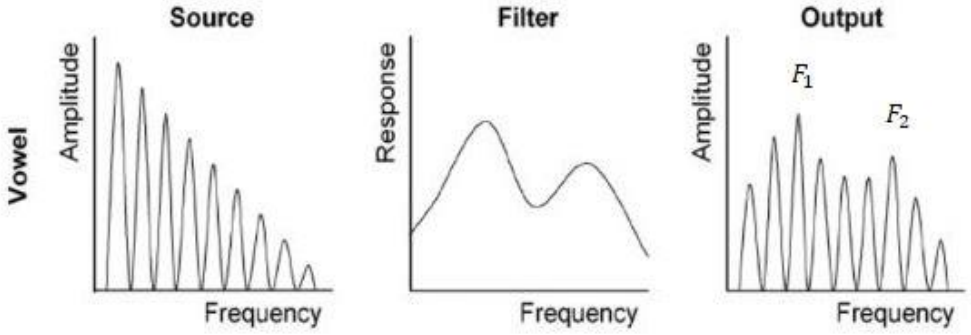


Figure 1.4: Filter effect of the vocal tract.

1.3 Overview of dysphonia

Humans rely on their voice to inform, persuade and interact with other people. For this reason, a phonation disorder can be very limiting, interfering with the ability to communicate and with normal daily activities; it has a negative impact on human life from a physical, social, emotional and economic point of view. People develop voice disorders for a variety of causes, from the improper use of vocal cords or allergies, to laryngeal cancer. Between these extremes there are many clinical cases responsible for dysphonia, which is the term generically used to indicate all possible vocal diseases, and these have to be conveniently diagnosed and treated. Therefore, the several laryngeal disorders that determine a total or partial phonatory inability have a wide variability: from simple hoarseness up to the aphonia, that is the total loss of voice. For these reasons, in recent years, many different analysis methods have been developing, both qualitative and quantitative, in order to improve the voice quality assessment.

"Dysphonia" is the medical term used to indicate a generic alteration of the voice, that can be qualitative and/or quantitative, temporary or permanent, of structural

origin or linked to one or more than one functional organs involved in phonation. At the level of perceptual analysis, this anomaly can be understood mainly as difficulty in controlling the pitch, the strain, the loudness or the voice quality. In these terms, v*oice quality* has been defined as variation of the overall timbre of a sound [2]; the latter is "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" [3], as stated by the American National Standards Institute. In addition, dysphonia can be associated with pain or discomfort while speaking.

The entire larynx is involved in the phonation, since its walls vibrate, producing a complex sound. The amplification and resonance phenomena occur thanks to the action of larynx, oral cavity, nasal cavities and paranasal sinuses. Finally, the emission of articulated language depends on voluntary movements of tongue, teeth, lips and cheeks. The damage to any of these structures can lead to an alteration in the production or control of the voice. For example, if the vocal cords become inflamed, develop protuberances or become paralyzed, they may not function correctly, causing dysphonia.

It is very difficult to classify clearly the several pathologies affecting the vocal folds; however, generally, voice disorders are divided into two groups: organic dysphonia and functional dysphonia. The former includes morphological or neuromuscular alterations, which can be related to:

- Laryngitis, acute or chronic; it causes a raspy or hoarse voice due to inflammation of the vocal cords.
- Congenital malformations.
- Neoplasia affecting an organ of the vocal apparatus (precancerous: dysplasia).
- Laryngeal trauma, linked to post-surgery results or accidental.
- Metabolic or endocrine diseases.
- Neurological diseases, such as Parkinson's.

By contrast, functional dysphonia consists in an excess or in a defect of the phonatory function. This type of dysphonia does not exhibit injuries and can be:

- Psychogenic, linked to muscular-tensive alterations in the larynx, weakening of the vocal apparatus muscles, functional alterations of the respiratory system or of psychic-emotional nature.
- Caused by vocal abuse.
- Idiopathic, without an apparent cause.

In detail, some of these pathologies, both organic and functional, could lead to the formation of noncancerous growths on the vocal cords. They can include nodules, polyps, and cysts; all this lesions cause the voice to be hoarse, low, and breathy and result typically from vocal overuse or vocal cord trauma.

Vocal nodules (fig. 1.5) are often a problem for professional singers. These small growths develop in the midpoint of the vocal folds; they look like calluses under the microscope and are occasionally associated with abnormal blood vessels. They most often grow in pairs, one on each cord. Generally, they form on parts of the vocal folds that undergo the most pressure when the cords come together and vibrate.

Figure 1.5: Vocal nodules.

Vocal cord polyps (fig. 1.6) are different from nodules; firstly, they can occur on one of the vocal folds. They tend to be more vascularized than nodules, meaning they have more blood vessels and appear reddish in colour. These growths can vary in size and shape, but are usually larger than nodules and are similar to blisters.

Figure 1.6: Vocal cord polyp.

Vocal cord cysts (fig. 1.7) are growths that have a sac around a fluid-filled or semisolid centre. They are less common than vocal cord nodules and polyps. There are two types of vocal cord cysts, mucus retention cysts and epidermoid (or sebaceous) cysts.

Figure 1.7: Vocal cord cysts.

Also a simple edema can afflict the vocal folds: it is a swelling (edema) of a vocal cord portion close to the edge, due to the whey, the watery part of the blood. Also in this case the vocal cords no longer have a good contact during the vibration and the voice begins to sound dirty, breathy, raspy, with a progressive reduction of the timbre, which becomes more opaque and brittle.

Others disorders that could affect specifically the vocal folds are hyposthenia and paralysis. The former, consists in the structural degradation of the vocal folds resulting in the reduction of muscle strength; the latter, happens when one or both vocal cords does not open or close properly. When one or both vocal cords are paralyzed, food or liquids can slip into the trachea and lungs; it is a serious disorder, since it can affect speaking, breathing and swallowing. In most cases, it is a consequence of a lesion that damages the innervation of the vocal cords. However, there are conditions in which the nerves are not damaged but are affected by inflammation. Finally, there is spasmodic dysphonia that appears with intermittent spasms of vocal fold muscles and its causes are unknown.

There is a wide range of causes that can make worse the vocal health status. Some of these are easy to evaluate and treat, while others require more attention, especially when they do not tend to improve over time or with standard therapies. The causes of voice dysfunctions may include upper respiratory infections, inflammation caused by acid reflux, improper use and vocal abuse, laryngeal nodules or papillomatosis, neuromuscular diseases (such as spasmodic dysphonia or paralysis of the vocal cords) and psychogenic conditions due to psychological trauma. There are many other factors that can be involved in the appearance or deterioration of dysphonia, such as gender, age, smoke, environmental conditions, stress condition, daily activities, weather conditions, etc.

In conclusion, it is important to remember that voice disorders are mostly reversible and can be cured successfully if diagnosed in time. Anyone can develop a dysphonia, but some professions are more susceptible: singers, actors, teachers, call-center employees, doctors, lawyers, nurses, sales people, public speakers, etc. Therefore, it is evident that nowadays a device that is able to measure acoustic parameters from vocal signal has been becoming increasingly necessary for an objective evaluation of the vocal health status. From this point of view, the main goal is to succeed in classifying healthy and pathological subjects with a high degree of reliability. Thus, the first part of this thesis deals with the research of parameters that can help with dysphonia recognition.

1.4 Vocal load and vocal effort in professional voice users

People who use their voice professionally for public speaking or singing often are subjected to voice disorders manifesting as hoarseness or breathiness, lowered vocal pitch, vocal fatigue, non-productive cough, persistent throat clearing, or throat ache. These symptoms often are related to benign lesions, such as vocal nodules, vocal fold edema or polyps. Such pathologies are usually caused by vocal fold hyperfunction, which is the excessive laryngeal muscular tension when speaking. Vocal fold tissue reacts to mechanical stress connected to abusive patterns of vocal behaviour, making voice disorders chronic.
About one third of the labour force have occupations in which the voice is the main tool [4] and it is likely that voice disorders develop in those individuals who use high-voice at work more than in others, as revealed by some studies [5, 6]. Existing literature uses to describe vocal fatigue through the concept of *vocal load* and *vocal effort*.

Vocal load is a combination of prolonged voice use and additional factors, such as elevated phonation frequency and high sound pressure level [7, 8, 9]. As suggested by its definition, it is assessed through the estimation of three acoustic parameters from the vocal signal:

- Voice Sound Pressure Level (SPL), at affixed distance (in dB).
- Fundamental frequency (F_0).
- Vocal dose.

There are different types of vocal dose, as explained by Titze, Sˆvec, Popolo et al. [10, 11], but in this work only the voicing time percentage (Dt%) is considered, that is the percentage of time spent phonating for the total monitoring period.

Moreover, intensively speaking, as any other demanding physiological voluntary activity, needs a certain effort. **Vocal effort** is defined as a physiological magnitude that takes into account changes in voice production, namely in vocal loading, caused by the distance from the listeners, noise and the physical

environment [12]. As a consequence, in order to evaluate the vocal effort, not only vocal load parameters, but also noise level measurements are needed. In detail, the background noise is commonly measured and expressed in terms of L_{A90} distribution. L_{A90} is a statistical parameter that is representative of the background noise level; it is the A-weighted noise level that is exceeded for 90 per cent of the measurement period.

Thanks to in-field and long-term monitoring of voice of those professional classes that force vocal folds continuously for work reasons, in the last few years, it has been possible to investigate the variations of vocal load during working hours. Simultaneously, objective environmental measurements have been made in order to evaluate the vocal effort that could bring to laryngeal dysfunctions. In particular, the attention has been paid on teachers: they have proved to be one of the categories mostly affected by vocal disorders linked to job.
Therefore, the second objective of this study focuses on the estimation of acoustic parameters for the vocal load assessment in teachers. In addition, their relationship with background noise has been taken into account for the vocal effort evaluation.

2. State of art

2.1 Vocal health assessment

2.1.1 Instrumental and perceptual evaluation of voice

Currently, people suffering from dysphonia are examined in a phoniatric clinic, where the specialist carries out an instrumental and perceptual evaluation of the patient's voice in order to recognise and diagnose vocal fold pathologies.

Firstly, the videolaryngostroboscopy is used as the most important clinical tool for instrumental voice assessment. It is a videoendoscopy with stroboscopy, i.e. a camera that records images thanks to the insertion of a flexible or rigid fiberscope; the images are projected on a video in real-time. This technique allows observing directly the anatomy and physiology of larynx and vocal cords and the muscle involvement during phonation. Moreover, it uses a flashing light so as to examine the vibration of vocal folds and their opening and closing: it represents an important method for identifying voice problems. However, such instrumental examination is intrusive, real-time and can only be performed in clinics, where people have a different vocal behaviour from everyday life. In fact, as specified by Manfredi *et al.* [13], physicians should know how their patients' voices sound in daily life, so that they could identify defective patterns in free speech and try to modify them.

Figure 2.1: Example of a vocal fold image in laryngostroboscopy

As far as the qualitative analysis of voice is concerned, there is not a specific exam or an objective analysis that are able to define the dysphonia severity. The most widespread approach is the perceptual evaluation of the voice quality by experienced phoniatricians; it consists in the use of an auditory-perceptual rating scale. Different voice quality rating protocols have been introduced; one of them is the GRBAS scale, or GIRBAS scale, widely used in Japan, since it has been proposed by Hirano (1981). In recent years, this scale has been becoming the standard scale for speech therapists and phoniatricians, also in Europe, where it is not officially recognised. As highlighted by table 2.1, every letter of the acronym refers to a qualitative characteristic of the voice: Grade of dysphonia (G), Instability (I), Roughness (R), Breathiness (B), Asthenia (A) and Strain (S). For each one, listening to the patient's sustained vowel /a/, the clinician has to assign a degree, a number in the range 0-3, where zero identifies a healthy voice and three a seriously unhealthy voice.

It is necessary to take into account that the auditory perceptual assessments are subjective; in fact, they depend strictly on the clinician experience. They are affected also by other many factors, such as the environmental conditions, the dysphonia degree, the type of perceptual scale, etc. In spite of such limitations, the perceptual evaluation of the voice quality goes on being an essential instrument, providing a "universal" language between the physicians. However, these technics alone are incomplete and not so reliable. Consequently, in the last few years, objective methods, based on the acoustic characteristics of the vocal signal, have

been spreading, since they attempt to quantify reliably the dysphonia severity. Therefore, this quantitative analysis would complete the clinician's diagnosis; other advantages are that it is non-invasive, relatively low cost and easy of application [14]. For this reason, the next two paragraphs deal with acoustic parameters that can be extracted from in-clinic recordings of voice.

GIRBAS scale	
Component	**Description**
G - Grade	General grade of dysphonia
I - Instability	Changes in voice quality over time
R- Roughness	Impression of irregularity of the vibration of the vocal folds
B - Breathiness	Degree of breath voice
A - Asthenia	Degree of weak voice
S - Strain	Degree of strain and hyperfunctional use of phonation

Table 2.1: GIRBAS scale.

2.1.2 Perturbation parameters and HNR

To by-pass the subjectivity that characterizes the auditory perceptual evaluation from experienced voice raters, many acoustic analysis algorithms and methods have been implemented: they provide an objective tool to assess voice problems, a numerical output that is easy to communicate to all interested people, such as phoniatricians, patients, third-party payers, and physicians. These techniques consist in the extraction of some acoustic parameters that can be seen as features of the vocal signal, in order to discriminate healthy and unhealthy voices. According to several studies, the fundamental frequency F_0 cannot be considered under this point of view, because it is influenced by gender, age, professional

uses, lifestyle, and many other factors. Consequently, the first investigated acoustic parameters were *jitter* and *shimmer*, measured in the time domain. As explained in chapter 1, from an acoustic point of view, the vocal signal is a complex and quasi-periodic sound; it exhibits more or less gradual variations in the fundamental frequency F_0 and in the amplitude. These variations, more properly defined as *perturbations,* which can be short-term or long-term, appear also in the normal voice production, within certain limits, making the human voice more natural. Short-term perturbations arise within few vibration cycles, sometimes between a cycle and the next. On the other hand, long-term perturbations involve a longer time, that includes many vibratory cycles, and they are at the base of the vocal tremor, physiological or not. The short-term perturbations of F_0 are defined jitter, while those of amplitude identify shimmer (fig. 2.2). In addition, jitter is linked to an uncontrolled vibration of the vocal folds; shimmer instead is associated with the presence of breathiness [15].

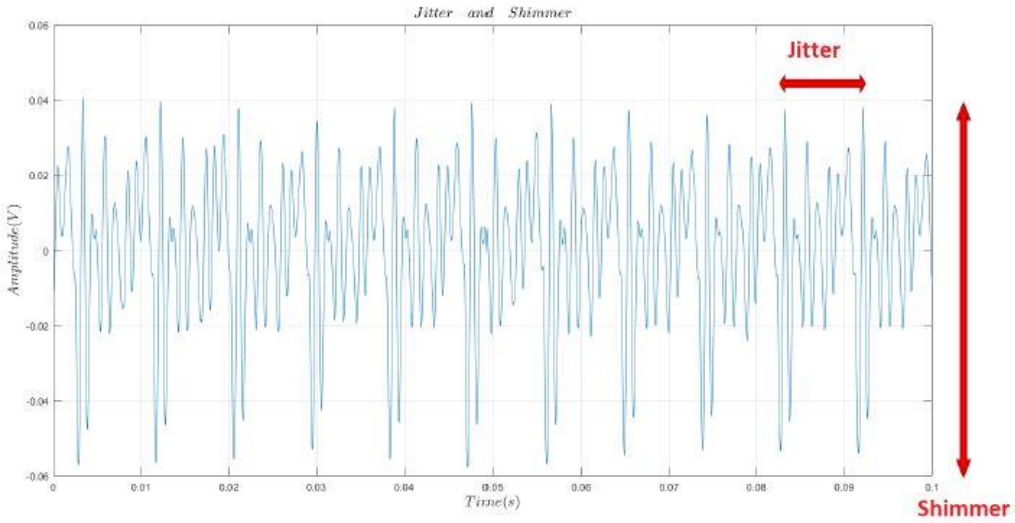


Figure 2.2: Jitter and shimmer in a vocal signal.

In the voice production, the presence of such irregularities over a certain physiological limit, leads to the periodicity dropping and to the noise overlay, making the vocal signal aperiodic; as a result, the voice is perceived as dysphonic. For this reason, many researches have studied effective, rapid and reliable systems to evaluate these perturbation parameters (jitter, shimmer and others connected to

them) for quantifying the irregularity of the vocal signal and correlating this to a voice dysfunction. Indeed, the perturbation level increases with the dysphonia severity.

Another feature of the vocal signal commonly used in acoustic analysis is the harmonic-to-noise ratio (HNR), which can be estimated in time domain or in frequency domain. The HNR expresses the ratio between the harmonic energy and the noise energy; in fact, as said before, in voice diseases, noise and harmonic structure overlap. It is the lower the higher is the aperiodic component of the vocal signal. This parameter was employed for the first time in 1982 by Yumoto et al., who worked in time domain [16]; but Qi and Hillman proved an easier computation in frequency domain [17].

All the above mentioned parameters must be extracted from continuous vowels produced with steady pitch and loudness, since any significant changes will be read as increments in vocal perturbation. However, this is not the only limitation that affects the use of traditional perturbation measures. In fact, as often highlighted in the existing literature, the great problem lies in the fact that they depend on the accurate detection of cycle boundaries, that is where a cycle of vocal fold vibration begins and ends, thus they become unreliable in extremely perturbed signals. In other words, from a computational point of view, the implementation of these parameters is based on the identification of the fundamental frequency: small errors in the estimation of F_0 afflict the measure of jitter, shimmer, HNR and all the other parameters connected to them [18]. As a result, they perceive as dysphonia any perturbation in the signal and do not seem to be good predictors of vocal disorders.

Despite such limitations, in the last few years, the F_0-based parameters have been reconsidered about the detection of voice pathologies. In 2006 P. Gomez-Vilda et al. demonstrated that F_0-based measures combined with biomechanical parameters increase the reliability of acoustic analysis, enhancing the identification of vocal fold lesions, such as polyps, nodules and Reinke's edema [19]. For example, as explained by Nicastri et al., the amplitude parameters are the best in the detection of polyps and cysts, because the not complete closure of the glottis creates an air escape that compromises the skill of producing a constant

sound emission (breathiness) and causes changes in the vibration amplitude without affecting the frequency [20].

2.1.3 Spectral and cepstral-based parameters

In order to overcome the limitations of perturbation parameters described in the previous paragraph, current practices are taking into account spectral- and cepstral-based measures, since they do not involve cycle boundary detection and can be got not only from sustained vowels but also from continuous speech that is able to stand for daily speaking patterns.

Many recent studies have been demonstrating the better accuracy of these new methods in detecting dysphonia: they allow achieving much more information of the vocal signal. Firstly, spectral analysis of digital vocal signals has been considered: it is based on the application of the Fast Fourier Transform algorithm (FFT) on sequential temporal windows of the signal analysed. Each spectrum, defined also *power spectrum*, gives information about the energy associated to the harmonic components (frequencies) of the waveform that corresponds to a specific time window. By combining in time domain all the spectra, it is possible to observe how the harmonic content of the vocal signal changes with time. The FFT spectra obtained in consecutive analysis windows can be averaged: the output is the *Avarage Power Spectrum* (APS), obtained using a short-time vocal window, or the *Long Time Avarage* Spectrum (LTAS), computed over long-time vocal simples. After that, some quantitative evaluations regarding spectral energy distribution can be performed; for instance, a greater concentration of energy at high frequencies, compared to medium-low ones, is considered indicative of hypofunctional vocal patterns. In 2011, Lowell et al. used the LTAS to recognise laryngeal pathologies. In detail, as highlighted by Lowell, in a healthy speaker, voice spectral energy expanded up to about 5 kHz; instead, dysphonic speakers had a wider frequency band, about 10 kHz [21].

Nevertheless, in recent years, many researches have been moving towards a new significant approach: the cepstral analysis. Developed in 1960, this technique is

able to extract accurately and automatically the fundamental frequency, by separating the glottal source patterns from the resonance characteristics of the vocal tract. By means of the vocal waveform FFT, a representation in the frequency-domain is obtained starting from one in the time-domain. With another application of FFT, but now on the power spectrum, a representation in the time-domain is achieved again (inverse FFT). Therefore, *cepstrum* is the spectrum of a spectrum; in the specific, it is the "log-power spectrum of the log-power spectrum of a signal", as stated by Borget et al. [22]. Since the cepstrum is a graph in the time-domain like the original vocal signal, new terms were invented to distinguish them, commonly used even now: *cepstrum* and *quefrency*, which are the inversion of spectrum and frequency respectively. Fig. 2.3 illustrates an example of spectrum and cepstrum of a vocal signal *y*.

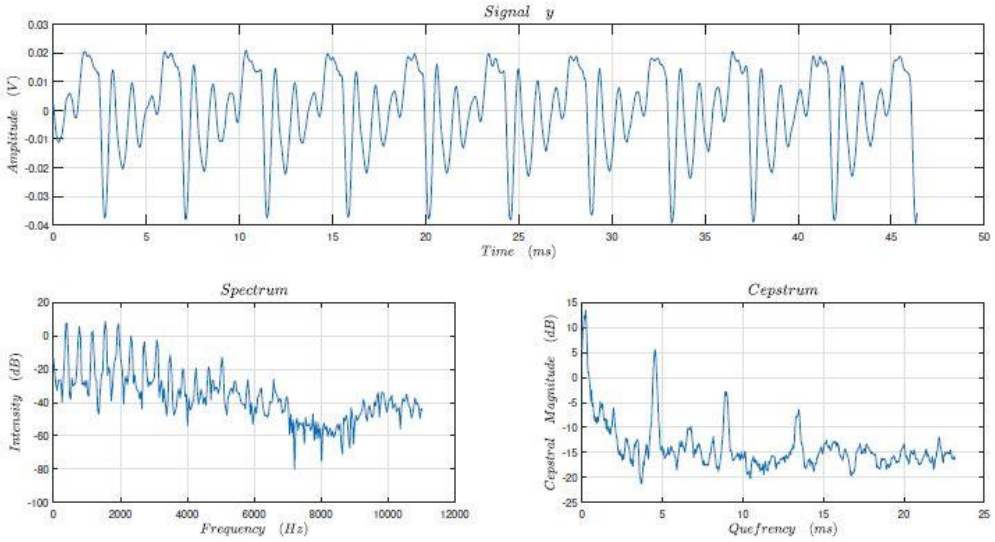


Figure 2.3: Spectrum and cepstrum of a vocal signal y.

Recently, the objective analysis of dysphonia has been focusing on the quefrency domain. In 1994, Hillenbrand et al. were the first in making cepstral measurements useful to predict voice dysfunctions, such as breathiness [23]. This is possible because the cepstral analysis, in addition to provide the F_0 value, allows for the assessment of the periodicity degree of the vocal signal. In fact, while the spectrum displays the frequency distribution of the signal energy, the cepstrum suggests how periodic the harmonic components in the spectrum are. Surely, in the cepstrum of periodic signals that have a well-defined

harmonic structure, a peak corresponding to fundamental period is highly visible (fig. 2.4); it is placed approximately between 3ms and 16ms (60 Hz and 300 Hz respectively). On the contrary, in dysphonic voices, this peak is lower and not so detectable.

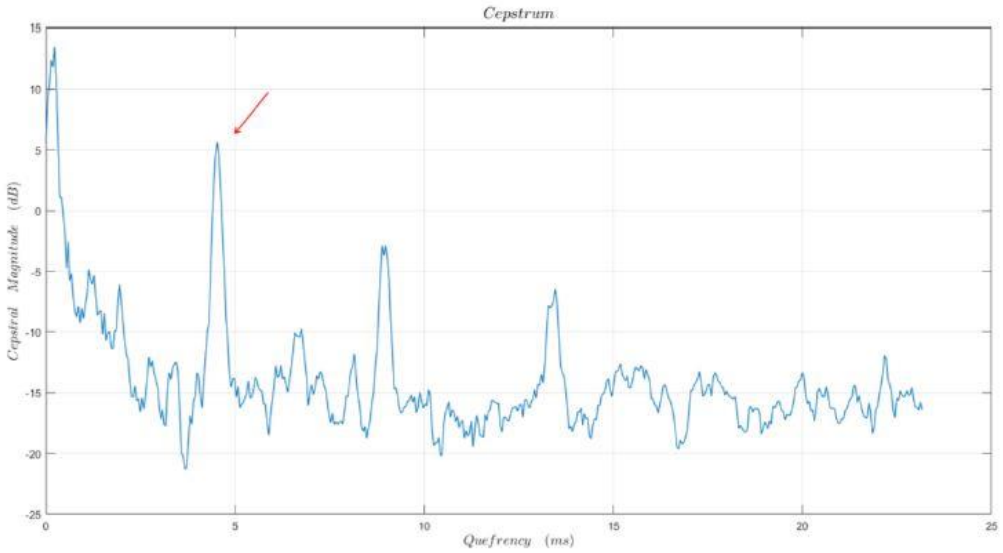


Figure 2.4: Cepstral peak.

Moreover, the absolute value of the peak amplitude (in dB) identifies the parameter *Cepstral Peak* (CP). However, the latter is affected not only by the harmonic organization degree, but also by overall signal energy and especially noise. Accordingly, the peak value in relation to background noise is more significant than the absolute value. For this reason, Hillenbrand, in his works, defined two cepstral parameters: the *Cepstral Peak Prominence* (CPP) and the *Cepstral Peak Prominence Smoothed* (CPPS). The former is calculated as the difference between the peak amplitude and the value at the same quefrency located in a linear regression line, which is fitted relating quefrency to cepstral magnitude. With the aim of having an improvement in prediction accuracy, Hillenbrand re-offered the same work introducing a modification in the CPP algorithm: two smoothing steps were considered before computing the normalized cepstral peak. Therefore, the CPPS is defined as the CPP, as suggested by fig. 2.5, but the cepstral magnitude derives firstly from an average across time of a few consecutive cepstra, and secondly from an average in the quefrency domain [24].

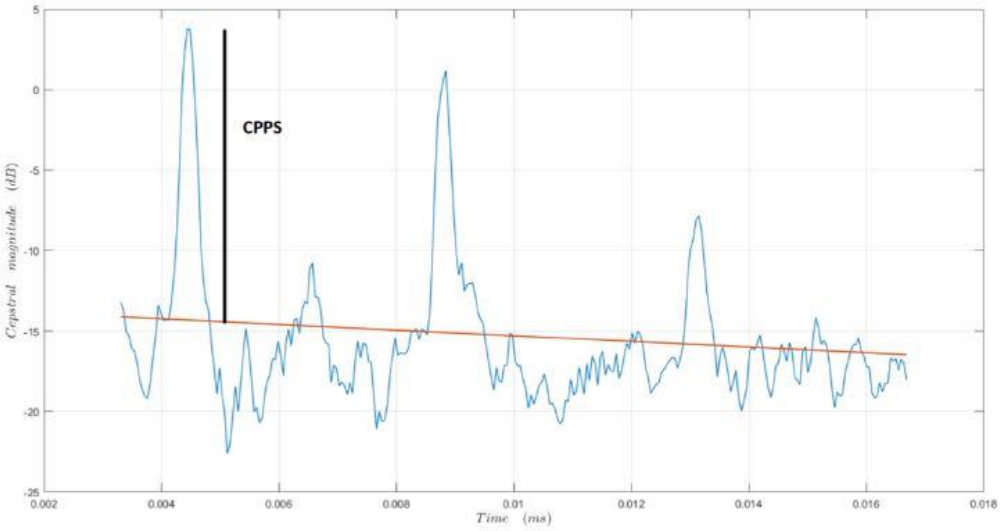


Figure 2.5: CPPS detection.

This parameter has acquired increasing relevance over time; in fact, it has been proved that it satisfies the meta-analysis on the correlation with perceptual evaluation of voice. In detail, CPPS resulted well correlated with the overall grade of dysphonia (G) and different types of voice quality in both sustained vowels and continuous speech.

Commonly, the CPPS is not calculated for a single cepstrum, but a window shifts along the time signal with a fixed overlap. The CPPS is defined for each frame so that the algorithm output consists of several CPPS values. Finally, it is possible to create a CPPS distribution and extrapolate different descriptive statistics. The first distribution value used by literature was the mean. However, in 2002, other researches went on studying breathiness and roughness considering not only the mean value Of CPPS distribution, but also others statistics, such as median, standard deviation and range [18]. The strong spread of studies and publications led to the birth of different software, which permitted to compute the mean parameters of a vocal signal; some of these tools are Praat, SpeechTool and ADSV. Consequently, it needed to identify differences and equalities of their algorithms. For example, in 2014 Maryn et al. compared Praat and SpeechTool results [25], and in 2016 Watt et al. made a comparison between Praat and ADSV that was reconsidered in a 2017 publication of Sauder et al. [26]. Furthermore, in

2017 Castellana et al. conducted important investigations in the voice analysis field: they focused on several descriptive statistics of the CPPS distribution as possible discriminators of vocal health in sustained vowel /a/. From their publication a relevant innovation emerged regarding the vocal signal acquisition; it is the use not only of a microphone in air, but also of a contact condenser microphone [27]. In another paper of the same researchers published in 2018, the fifth percentile of the CPPS distribution turned out to be the best discriminant between healthy and unhealthy voices in a sustained vowel /a/ acquired with a microphone in air, while the standard deviation was found for the contact microphone [28]. Another 2018 study of Castellana et al. shows the results about the ability of the CPPS to discriminate pathological and healthy subjects in continuous speech: the ninety-fifth percentile was the best diagnostic parameter in both reading and free speech [29].

The latter three cited publications are the starting point for the first part of this thesis, whose target is the validation of the classification models previously achieved and the research of new ones for both sustained vowel and continuous speech. In order to obtain major information, different technologies of microphone have been examined: microphone in air, electret condenser microphone and piezoelectric contact microphone. Particular attention has been payed to the results provided by other parameters in sustained vowel and the combined use of two variables for the research of the best classifier.

2.2 Vocal load assessment in teachers

As specified in paragraph 1.4, different levels of vocal disorders have been damaging those occupational categories that make use of voice in a sustained way and for long periods of time, such as actors, singers, call center employees, sales people, etc. The arising of voice problems could lead to absenteeism from work for recovering, with a resulting impact on the economy in terms of health care use, voice-related and loss of productivity at work [30].

For this reason, many researches have been conducted about the correlation between vocal loading parameters and L_{A90} background noise levels, thanks to long-term voice monitorings of occupational voice users during their working activities. Portable vocal analysers were used for this purpose [31]; they can be seen as voice dosimetry devices which measure the speech SPL of the speaker at a fixed distance (in dB), the F_0 (in HZ) and the Dt% (in %). The outcomes about F_0 and SPL are usually illustrated as histograms of occurrences that allow observing relevant features of the speaker vocal behaviour over many hours. Therefore, it was demonstrated the involuntary tendency of speakers to raise their voice level as the noise level rises for enhancing intelligibility of the speech signal; this phenomenon is called *Lombard effect* [32, 33]. All in all, by revealing the changes of vocal loading during working hours and their relationship with the background noise levels, in recent years, it has become possible to quantify vocal effort and so identify the risk of vocal dysfunctions [34].

One of the largest categories of professional voice users are teachers. Teacher's voice is susceptible to disorders resulting from prolonged voice use and heavy vocally loading conditions [35]. Several authors have investigated the high prevalence of recurring symptoms of vocal overloading and fatigue in teachers. In particular, according to a statistical study, the correlation between the occurrence of voice disorders and occupational voice use has been observed in the 58% of cases for teachers and in the 29% of cases for other occupational voice users [36]. Lack of voice training, unawareness of suitable vocal hygiene and poor environmental and working conditions, all may be responsible for the development of voice dysfunctions in teachers. For instance, in presence of adverse environmental conditions, such as background noise, teachers are inclined to raise their voice and speak with higher vocal loudness resulting in a risen vocal effort and strain [35]. In literature, several studies have been carried out using a portable voice analyser: SPL, F0, and the phonation time were estimated and they were significantly greater in teaching conditions compared with non-teaching conditions. …….Such a result underlines a risky situation for teachers at work. Accounting for these observations, the second objective of this thesis is the vocal

load assessment of primary school teachers by means of in-field voice monitoring, and its relationship with background noise levels measured with a specific device, which is able to return a visual feedback related to noise level.

3 Acoustic parameters as predictors of voice health status

3.1 Data collection

3.1.1 Subjects

In this study, data collection has been achieved through voice recordings of two subject groups: healthy and unhealthy subjects. Regarding dysphonic voices, 102 voluntary patients have participated. Some of them correspond to the dataset described in [27] and the respective data have been collected between 2015 and 2016. While the recordings related to 36 of 77 unhealthy subjects have been taken between May and September 2017, and the remaining 25 ones between April and June 2018. All the 77 patients are native Italian speakers and suffer from different vocal diseases. This group consists of 73 females and 29 males, with an age range from 20 to 82 years old (mean age: 54.6 years and standard deviation: 18.0 years). Table 3.1 summarizes the diagnosis for the patient group. On the same time, other 73 voluntary subjects, 36 females and 37 males, have been involved in the

research to form the control group. The latter is essential for creating a healthy-unhealthy classifier because it includes data from normal subjects that are taken as reference. In this case, it is made up of people with healthy voices, who are also native Italian speakers with age between 19 and 58 years old (mean age: 26 years and standard deviation: 6.3 years). It is important to underline that, in a previous work, some results about healthy-unhealthy discrimination were found using much of the above mentioned database; in detail, 77 patients and 64 controls formed the *training set*. For this reason, in some experiments of this study, the remaining 25 patients and 9 healthy subjects have made up the *test set,* on which the models found earlier have been applied in order to evaluate their generalization ability. Differently, in other experiments, the whole data base has been used in order to search for new classification models.

Type of dysphonia	**Number of patients**
Cyst	16
Edema	14
Sulcus vocalis	6
Polyp	6
Chronic laryngitis	9
Vocal fold hypostenia	10
Vocal fold paralysis	12
Vocal fold nodul	6
Neurological disorder	6
Post-surgery dysphonia	3
Spasmodic dysphonia	1
Functional dysphonia	13
Overall	102

Table 3.1: Types of dysphonia in patients of the dataset.

All the voice recordings were performed in an ambulatory room of otolaryngology department of San Giovanni Battisti Hospital, in Turin, during phoniatric examinations. The room was often exposed to ordinary noise that could have

affected some vocal signals; for this reason, inside the room, during the recording procedure, silence should have been done.

3.1.2 Procedure

The recording procedure was conducted in the phoniatrics department at the end of a medical examination. The latter, consists of several steps that allow the physician to define the patient voice as healthy or pathological. Firstly, it is based on previous reports, on the patient clinical history and includes the auditory-perceptual evaluation of voice quality through the assignment of a GIRBAS rating. After that, the phoniatrician carries out the videolaryngostroboscopy for investigating the health status of the vocal apparatus: using a flexible or rigid fiberscope, provided with flashing light, he can observe the vocal folds, their movements of opening and closing, the muscle involved. In this way, he can establish the type of dysphonia in pathological patients.
For engineering purposes, another step was added to the medical examination: the acquisition of voice samples, the real important part for this work. During this step, the subject wore three different microphones, two contact and one in air (fig. 3.1). From this point on, the voice recording phase began; in detail, each volunteer was asked to follow a conventional protocol made up of three tasks:

a. Vocalize the vowel /a/ three times, on a comfortable pitch and loudness. If possible, each vowel has to be maintained from 3 to 10 seconds. Between a vowel and the next, the subject can wait for the necessary for catching his breath.

b. Read a phonetically balanced Italian passage (Appendix A), without interruption from the beginning to the end; this task is commonly called “Reading”.

c. Speak for one minute, freely and with no stop; this is the “Free speech” task. In this case, the topic is not important in order to capture all the nuances of the voice when it is not influenced to specific patterns.

After the recordings, the participants filled in a questionnaire, called “Profilo di Attività e Participazione Vocale” (PAPV). As shown by fig. 3.2, it is composed by 28 questions, divided into five sections that are about a self-evaluation of the dysphonia severity and the effects perceived on different situations of daily life, such as job, social activities, relationship with relatives, friends, collegues, etc. [37].

It is important to underline that, during data collection, for each subject different information about their own were gathered: age, gender, vocal disease, job, smoker or no-smoker, GIRBAS scale and the valuator. Fig. 3.1 gives the idea of the recording environment and shows examples of subjects wearing the three microphones.

Figure 3.1: Recording environment in clinic.

NOME: COGNOME:

ETA': SESSO: M F

FUMATORE: SI NO NOTE:

PROFILO DI ATTIVITÀ E PARTECIPAZIONE VOCALE - PAPV
Fava, Pontillo, Oliveira, Behlau

Ti invitiamo a rispondere ad ogni domanda mettendo una croce (X) su qualsiasi punto della linea che rappresenta al meglio il grado della tua risposta. Una croce sull'estremo lato sinistro della linea indica che il problema non è MAI presente; una croce sull'estremo lato destro della linea indica che il problema è GRAVE o è SEMPRE presente; una croce su qualsiasi punto della linea tra i due estremi, andando da sinistra a destra, indica che il problema è gradualmente più grave o più frequente

Autopercezione dell'entità del problema vocale

1 Attualmente qual è l'entità del tuo problema vocale?
Lieve — Grave

Effetti sul lavoro

2 **Il tuo lavoro risente del tuo problema vocale?**
Mai — Sempre

3 Negli ultimi 6 mesi hai pensato di cambiare lavoro a causa dei tuoi problemi vocali?
Mai — Sempre

4 Il tuo problema vocale ha creato condizioni di stress sul tuo lavoro?
Mai — Sempre

5 Negli ultimi 6 mesi il tuo problema vocale ha influito sulle decisioni legate al futuro della tua carriera?
Mai — Sempre

Effetti sulla comunicazione quotidiana

6 A causa del tuo problema vocale la gente ti chiede di ripetere ciò che hai appena detto?
Mai — Sempre

7 Negli ultimi 6 mesi hai mai evitato di parlare con gli altri a causa del tuo problema vocale?
Mai — Sempre

8 La gente ha difficoltà a capirti al telefono a causa del tuo problema vocale?
Mai — Sempre

9 Negli ultimi 6 mesi hai ridotto l'uso del telefono a causa del tuo problema vocale?
Mai — Sempre

10 Il tuo problema vocale influenza il tuo modo di comunicare in ambienti silenziosi?
Mai — Sempre

11 Negli ultimi 6 mesi hai mai evitato conversazioni in ambienti silenziosi a causa del tuo problema vocale?
Mai — Sempre

12 Il tuo problema vocale influenza il tuo modo di comunicare in ambienti rumorosi?
Mai — Sempre

13 Negli ultimi 6 mesi hai mai evitato conversazioni in ambienti rumorosi a causa del tuo problema vocale?
Mai — Sempre

14 Il tuo problema vocale influisce su ciò che vuoi comunicare quando parli a un gruppo di persone?
Mai — Sempre

15 Negli ultimi 6 mesi hai mai evitato conversazioni di gruppo a causa del tuo problema vocale?
Mai — Sempre

16 Il tuo problema vocale ti impedisce di far capire quello che vuoi comunicare?
Mai — Sempre

17 Negli ultimi 6 mesi hai mai evitato di parlare a causa del tuo problema vocale?
Mai — Sempre

Effetti sulla comunicazione sociale

18 Il tuo problema vocale influisce sulle tue attività sociali?
Mai — Sempre

19 Negli ultimi 6 mesi hai mai evitato attività sociali a causa del tuo problema vocale?
Mai — Sempre

20 I tuoi familiari, amici o colleghi di lavoro sono infastiditi dal tuo problema vocale?
Mai — Sempre

21 Negli ultimi 6 mesi hai mai evitato di comunicare con i tuoi familiari, amici o colleghi di lavoro a causa del tuo problema vocale?
Mai — Sempre

Effetti sulle tue emozioni

22 Sei infastidito dal tuo problema vocale?
Mai — Sempre

23 Ti vergogni del tuo problema vocale?
Mai — Sempre

24 Hai poca stima di te stesso a causa del tuo problema vocale?
Mai — Sempre

25 Sei preoccupato per il tuo problema vocale?
Mai — Sempre

26 Ti senti insoddisfatto a causa del tuo problema vocale?
Mai — Sempre

27 Il tuo problema vocale influisce sulla tua personalità?
Mai — Sempre

28 Il tuo problema vocale incide sulla tua immagine?
Mai — Sempre

DATA:

Figure 3.2: PAPV questionnaire.

3.1.3 Recording equipment

During recording step, voluntary subjects wore three different microphones: a microphone in air and two contact microphones. The former records the vocal signal after its passing through the vocal tract, which acts as a filter; as a result, the obtained signal is more complex. Moreover, this microphone acquires also the external noise, which overlaps the useful signal. With the aim of making comparisons between devices with different characteristics and evaluating the device dependence of results, the other two contact microphones have been employed: they are able to detect the vocal fold vibrations, providing as output a signal that is more similar to the glottis one. Moreover, the signal acquired by the two contact microphones are affected by a negligible background noise. However, the recording is less clear to the ear than the one obtained with microphone in air.

The sensors are:

- An omni-directional ***headworn microphone*** MIPRO MU-55HN (fig.). The sensitive element presents a flatness of ±3 dB in the range between 40 Hz and 20 kHz. It is connected to a bodypack transmitter ACT-30T, which transmits the signal to a wireless system Mipro ACT 311. A recorder ZOOM H1 (Zoom Corp., Tohyo, Japan) captures the output signal of the wireless system and stores it in a SD card with a sampling rate of 44100 Hz and 16 bit of resolution. The microphone is placed at a distance of about 2,5 cm from the talker lips.

Figure 3.3: Headworn microphone (MIPRO) and its transmitter.

- A contact ***Electret Condenser Microphone*** (ECM AE38, Alan Electronics GMbH (Dreieich, Germany)) (fig.). It is positioned on the jugular notch and fixed with a surgical band, so that the vibrations of vocal folds can be detected through the skin movements. The sensing unit is connected to a recorder ROLAND R05 (Roland Corp., Milano, Italy) that records the signal sampling it at 44100 Hz with 16 bit of resolution and stores it in a SD card..

Figure 3.4: Electret Condenser Microphone (ECM).

- A contact ***Piezoelectric Contact Microphone*** (HX-505-1-1, HKKK, 406, PLant 1, Jiadind Science Park, Dalang, Longhua New Dist., Shenzhen, Guangdong, China) (fig.). It is a neck-ring, whose sensing element has to be placed near the jugular notch. Also this sensor is sensitive to vocal fold

vibrations, but it is connected to a smartphone (Samsung SM-G310Hn) provided with the "Vocal Holter" App. that allows to record the signals with a sampling rate of 22050 Hz and a resolution of 16 bit.

Figure 3.5: Piezoelectric Contact Microphone.

Table 3.2 displays the details related to the subjects who undertook the experimental voice tasks with the three microphones. It is evident the reduced dataset, both patients and controls, used for vocal signals acquired with piezoelectric microphone: their analyses had been included later compared to the other devices.

	MIPRO MU-55HN			ECM AE38			PIEZO HX-505-1-1		
	M	F	Overall	M	F	Overall	M	F	Overall
Patients	29	71	100	23	63	86	14	37	51
Controls	28	25	53	36	29	65	16	19	35
Overall	57	96	153	59	92	151	30	56	86

Table 3.2: Number of participants to the experiments with the different devices.

3.1.4 Data pre-processing

After the voice samples recording, data was downloaded from the three microphones and saved in a Personal Computer as audio ".wav" files, through SD cards. In order to simplify the next processing, the signals recorded with ECM and MIPRO microphones were resampled at 22050 Hz, while the PIEZO signals were maintained at the original sampling rate (22050 Hz). After that, each recording was renamed according to the patient's identification code and cut into five different audio files using the Audacity 2.2.2 software:

- *Vowel /a/ files*: they are three for each recording (A1, A2 and A3) and around 10 seconds long, with the exception of some unhealthy subjects who did not manage to maintain the vowel for long time. Generally, they were got by chosing the 4 central seconds of the vowels, where the signal is more stable, excluding the initial and final part.

- *Reading file*: it was obtained selecting the reading part related to the "Bulka" passage from the original audio file. In particular, only the first nine sentences, about 120 words, were analysed, in order to reduce the computational time.

- *Free speech file*: it was obtained cutting only the first 30 seconds of the 1 minute recorded, for the same reason mentioned above.

Once all data were collected, selected and organised, the next step was the processing of the signals: they were submitted to several algorithms. In particular, for all the three types of audio file, the CPPS algorithm has been executed in order to estimate the CPPS distribution and calculate its descriptive statistics.

3.2 Methods and data processing

3.2.1 Perturbation parameters and Harmonic to Noise Ratio in sustained vowel

As explained in chapter 2, pathological voices are characterized by excessive perturbations in the fundamental frequency F_0, or in the fundamental period T_0, and in the amplitude of the vocal signal, which thus loses its periodicity because of noise overlap. Consequently, in addition to the analysis based on CPPS distribution, another method to discriminate a dysphonic voice from a healthy one could be the extraction of some parameters able to quantify the aperiodicity grade of the vocal signal. The main limitation is that they are valid only for sustained vowels, not in continuous speech. In addition, these perturbation parameters, according to their definition, are strictly related to the fundamental period, so that their implementation is deeply influenced by the T_0 identification. Part of this work consisted in the development of algorithms allowing for the estimation of such perturbation measurements and their application on the three sustained vowels /a/ of each acquisition. For this reason, the definition of these acoustic parameters is now provided. Firstly, they can be divided into two groups: parameters related to fundamental period perturbations and parameters related to amplitude perturbations.

Among the parameters that measure T_0 perturbations, were implemented:

- **Jita** (μs). It is the absolute jitter and describes the absolute mean variation period to period of the fundamental period T_0 (Ferrero et al.,1995 [38]):

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

where $T_0^{(i)}$, with $i = 1, 2,\ldots,N$, are the periods extracted by the vocal signal and N is the number of periods. This parameter describes the mean of the difference between one period and the next one.

- **Jitt** (*%*). It is the local jitter and describes the relative mean variation period to period of the fundamental period:

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

where $T_0^{(i)}$, with $i = 1, 2,\ldots,N$, are the periods extracted by the vocal signal and N is the number of periods. The formula is similar to the Jita one; it differs from Jita in the division for the average fundamental period.

- **RAP** (*%*). It is the Relative Average Perturbation of 3 in 3 periods with the step of one of the fundamental period. The formula is:

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

where $T_0^{(i)}$, with $i = 1, 2,\ldots,N$, are the periods extracted by the vocal signal and N is the number of periods. The RAP is similar to Jitt, but in this case, instead of calculating the difference between one period and the next, the average of three periods is calculated (3 as smoothing factor), then is subtracted the value of the central period.

- **PPQ** (*%*). It is the Pitch Period Perturbation Quotient and gives the relative average perturbation of 5 in 5 periods (5 as smoothing factor):

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

where $T_0^{(i)}$, with i = 1, 2,…,N, are the periods extracted by the vocal signal and N is the number of periods.

- **vF_0.** It is the Fundamental Frequency Variation. It is the relative variability of standard deviation of F_0 with respect to the calculated mean fundamental frequency:

$$vF_0 = \frac{\sigma}{F_0} \times 100 = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_0 - F_0^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N} f_0^{(i)}} \times 100$$

where F_0 is the average fundamental frequency, σ is the standard deviation of f_0, and $f_0^{(i)}$ are the individual frequency values extracted.

On the other hand, the parameters related to amplitude perturbations that have been considered in this study are:

- **ShdB** (*dB*). It is the absolute shimmer that describes the absolute average variability period by period of the peak to peak amplitude:

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

where $A^{(i)}$, with i = 1,2,..,N, are the amplitudes peak to peak and N is the number of impulses extracted. The absolute shimmer is very sensitive to the amplitude variations occurring between consecutive pitch periods, so it gives a measure of the short-term amplitude perturbation.

- **Shim** (*%*). It is the local shimmer and describes the relative evaluation of the period-to-period (very short term) variability of the peak-to-peak amplitude:

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

where $A^{(i)}$, with i = 1,2,..,N, are the amplitude peak to peak and N is the number of impulses extracted. Both Shim and ShbB are relative evaluations of the same kind of amplitude perturbation but they use different measures for the result, percent and dB.

- **APQ** (%). It is the Amplitude Perturbation Quotient and describes the relative variability of 11 to 11 periods (11 as smoothing factor) with step of 1:

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

where $A^{(i)}$, with i = 1,2,..,N, are the amplitudes peak to peak and N is the number of impulses extracted. APQ is less sensitive to pitch extraction errors than shim, but it still provides a reliable indication of short-term amplitude variability in the voice.

- **vAm** (%). It is the Peak Amplitude Variation. It gives relative variability of the peak-to-peak amplitude variations (short to long-term) within the analysed voice sample:

$$vAm = \frac{\sigma}{A_0} \times 100 = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_0 - A^{(i)})^2}}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}} \times 100$$

where $A^{(i)}$, with i = 1,2,..,N, are the amplitude peak to peak, A_0 is the average value of the extracted peak-to-peak amplitude and N is the number of impulses extracted.

Another acoustic parameter that was investigated by previous literature as a potential vocal health discriminator, is the Harmonic to Noise Ratio (**HNR**, dB). It is a measure that quantifies the amount of additive noise in the voice signal. It is the ratio between the components of harmonic spectral energy and the components of disharmonic spectral energy:

$$HNR = 10log_{10}\frac{r(\tau_{max})}{1 - r(\tau_{max})}$$

where $r(\tau_{max})$ is the local maximum of the normalized autocorrelation function and according to Boersma [39] it represents the relative power of the periodic (or harmonic) component of the signal, and its complement represents the relative power of the noise component.

3.2.2 Vowel /a/ processing

After the pre-processing step, each vowel /a/ file was processed in order to perform the feature extraction: two different Matlab® R2018a scripts were used for estimating several acoustic parameters, which could support the objective analysis of dysphonia and its severity.

The first algorithm was implemented for computing some of the perturbation parameters and the harmonic to noise ratio parameter, whose definition are exposed in above paragraph. The Matlab script operates on one signal at time and execute an operation of autocorrelation, in which the maximum index is extracted in order to find the fundamental frequency of the vocal signal and the corresponding pitch period. Starting from this fundamental period (T_0) in samples, the signal position is moved ahead by the current T_0 for computing all the T_0 for jitter related parameters and all peak-to-peak amplitude for shimmer related parameters. The HNR parameter is got calculating the fundamental frequency for signal windows of 1024 samples. The values obtained from all the windows are averaged to have a single value of HNR for each signal.

The second algorithm employed on vowels allowed to obtain the CPPS distribution and its descriptive statistics; the implementation is explained in section 3.2.4.

3.2.3 Continuous speech processing

Before applying the CPPS algorithm to Reading and Free speech signals, the latter have to be modified: it needs to remove those portions of the signal where there is not voice, namely the subject does not speak. A specific algorithm, the silence removing algorithm, has been used for this purpose: it succeeds in recognizing voiced and unvoiced segments of the signal. However, during the silence, the signal is not zero, but it has low amplitudes because of background noise, which probably modifies the final result of the next data processing. Specifically, the application of CPPS algorithm (explained in the next section) on silence segments leads to an alteration of CPPS distribution; additionally, the respective values of CPPS are not explicative of the subject condition.

The silence removing algorithm is implemented on software Matlab® R2018a. The script is based on finding a threshold, suitable for signal, between voiced and unvoiced segments. Firstly, the signal is divided into frames of length 1024 samples (46 ms). For each frame, the RMS value is calculated. Then, the mean value of all the RMS values is computed. The latter, is multiplied for *1/k*, where *k* is an "empirical" factor influenced by external noise. It is thus possible to adapt the threshold according to the amount of noise for each signal. Therefore, the multiplicative factor has been chosen equal to 1,9 for the signals caught by means of the microphone in air. Concerning the contact microphones, the value of the factor k is not so relevant, because they are not affected by external noises. The expression *RMSmean/k* defines the threshold. Subsequently, the RMS value of each frame is compared with the threshold: if the value is lower than the threshold, that frame is considered "unvoiced", and so it is rejected. By contrast, if the RMS value is higher than the threshold, the frame is maintained. Fig. offers an example of signal before and after the implementation of silence removing algorithm.

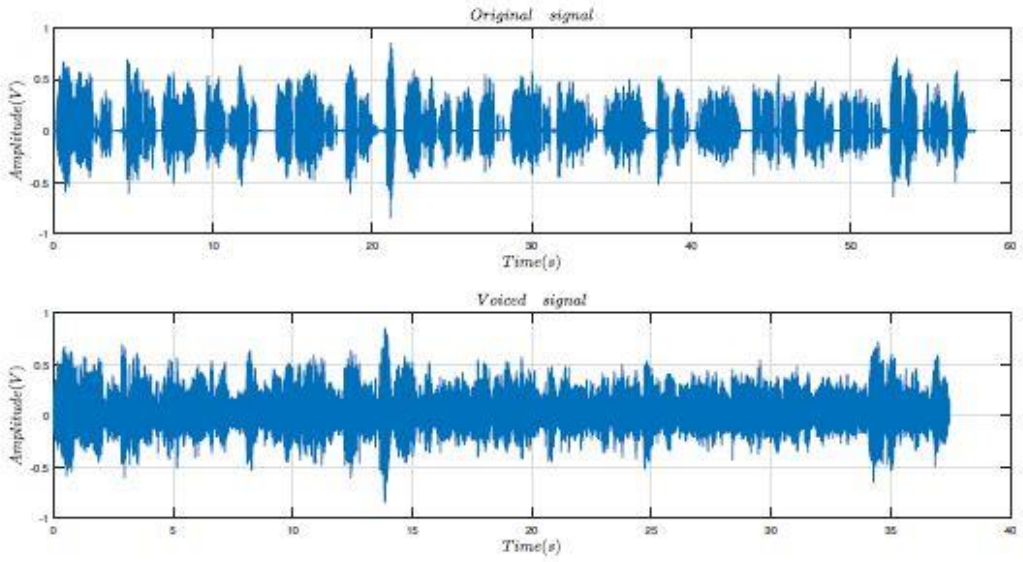


Figure 3.6: A vocal signal before and after the application of the silence removing algorithm.

Once Reading and Free speech signals contain only voiced samples, their processing can carry on with the CPPS algorithm, described in the next paragraph.

3.2.4 CPPS algorithm

After the pre-processing phase, the CPPS distribution was estimated for both sustained vowels and continuous speech signals, where the silence was previously removed. This analysis was performed on software Matlab® R2018a. As already said in chapter 2, the cepstrum is a log power spectrum of a log power spectrum [22]. Therefore, given the vocal signal *y*, it is possible to define:

$$yFFT = 20log\,|FFT(y)|$$

$$yFFT2 = 20log\,|FFT(yFFT)|$$

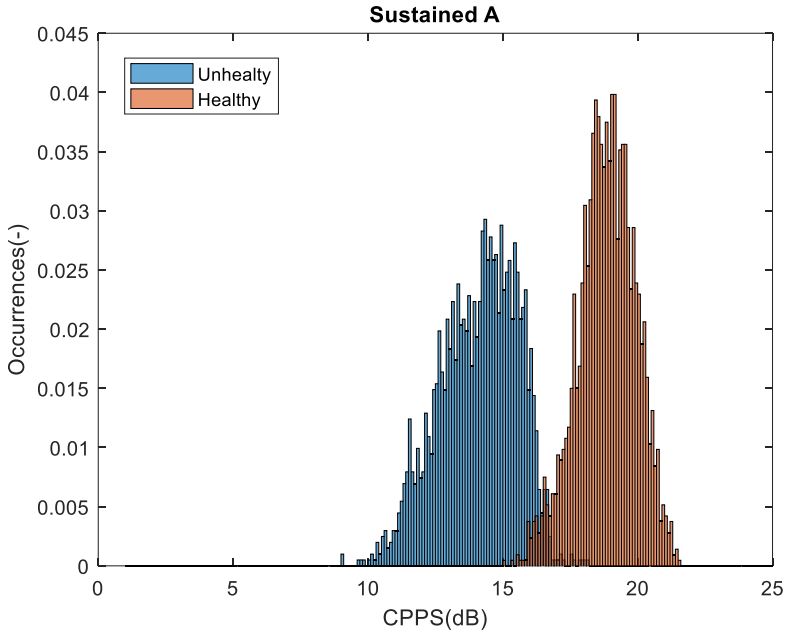
where *yFFT* is the spectrum of signal *y,* while *yFFT2* is the cepstrum..

The Cepstral Peak Prominence Smoothed is a measure (in dB) of the cepstral peak amplitude, normalized for overall signal amplitude through a linear regression line estimated relating quefrency to cepstral magnitude. In addition, it considers two smoothing steps, one across time and the other across quefrency (Hillenbrand et al., 1996). From a computational point of view, the CPPS algorithm divides the signal analysed into frames of length 1024 samples (about 46 ms), so that one frame at time is processed. The frames are overlapped, since the time window scans the complete signal with a translation of 2 ms. For each 1024-frame the spectrum is calculated with the Matlab function FFT (Fast Fourier Transform) and is multiplied by the Hamming window. After that, according to the formula, a second FFT is performed on the spectrum, leading to the cepstrum computation. This operation is repeated for each frame resulting in a certain number of cepstra. The cepstra of the frames are time averaged with a window of 14 ms (7 frame), then they are quefrency averaged with 7 bin windows. After the double smoothing, for each cepstrum the linear regression line is fitted: according to the definition, the CPPS is calculated as the distance between the cepstrum maximum peak and the corresponding value of the regression line, at the same quefrency. Specifically, the maximum search occurs between about 3.3 ms and 16.7 ms, because the range of fundamental frequency in human voice is from 60 Hz to 300 Hz. The final output is a distribution of CPPS, one for each frame; it can be plotted in the form of a histogram, with the bin size equal to 0,1 dB (fig.).

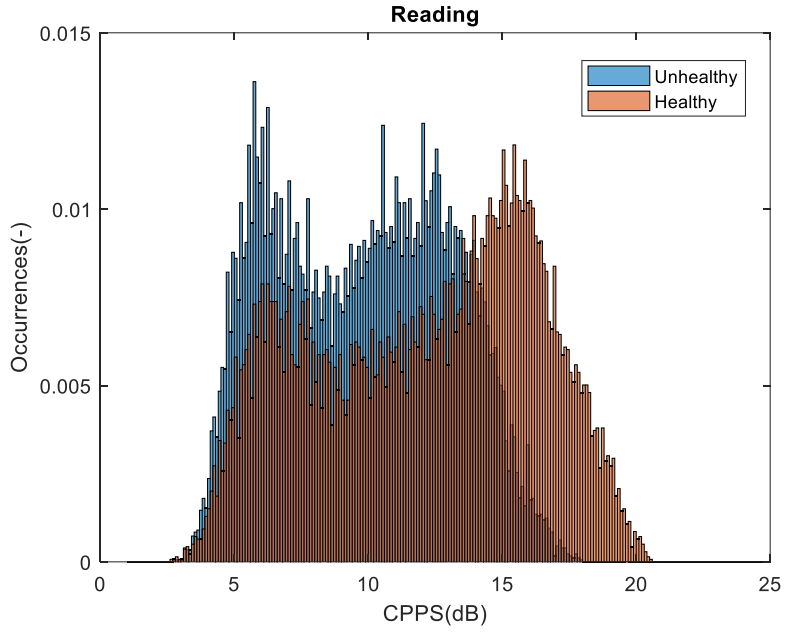
From the distribution, the following descriptive statistics are calculated: *mean* ($CPPS_{mean}$), *median* ($CPPS_{median}$), *mode* ($CPPS_{mode}$), 5^{th} *percentile* ($CPPS_{5prc}$), 95^{th} *percentile* ($CPPS_{95prc}$) as measures of location of the distribution; *standard deviation* ($CPPS_{std}$) *and range* ($CPPS_{range}$) as measures of its variance, *skewness* ($CPPS_{skewness}$), *and kurtosis* ($CPPS_{kurtosis}$) for the characterization of distribution shape. The next step is the investigation of their capability in discriminating between healthy and dysphonic voices, and this is possible by means of methods from statistical analysis.

In fig. 3.7, some examples of CPPS distributions belonging to a healthy subject and an unhealthy one are depicted for sustained vowel, Reading and Free speech.

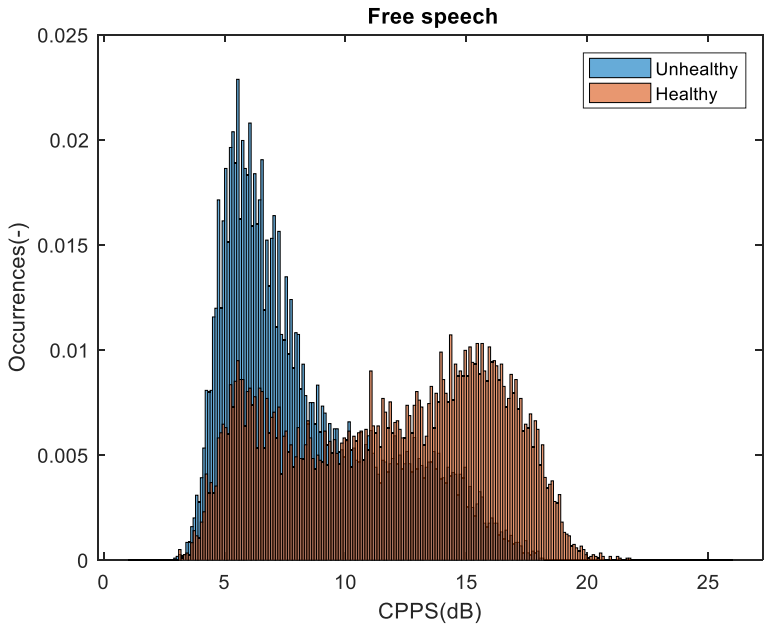
It is possible to observe that unhealthy voice shows a distribution with a lower mean (moved to the left compared to healthy voice).



a)



b)



c)

Figure 3.7: CPPS distributions of a healthy voice and an unhealthy one in sustained vowel (a), Reading (b) and Free speech (c).

3.2.5 Statistical analysis

By means of CPPS algorithm, the nine statistical parameters of CPPS distribution were extracted from all the different speech materials. Additionally, only the sustained vowels were submitted to the Matlab script that computes the perturbation measures and HNR, in order to make available other parameters in the vocal health investigation. Therefore, all the parameters were calculated for the 175 subjects (102 unhealthy and 73 healthy), starting from the original signals of the three microphones, where available. In fact, not all the subjects had the three recordings, because it needs to consider that the piezoelectric microphone was introduced at a later time; moreover, some signals were rejected or were not available for technical reasons (excessive noise, malfunction, low battery, etc.).

In order to be able to conduct a statistical study, for each microphone, a database was created; it includes identification code of subject, age, gender, grade G of

GIRBAS scale and the results of the feature extraction: the nine descriptive statistics of CPPS distribution and, only for vowels /a/, the perturbation parameters and HNR. In addition, with the aim of evaluating the effectiveness of the acoustic parameters as discriminators between healthy and unhealthy voices, a binary classification approach was followed. So, the value of another variable was assigned to each subject; it is equal to 0 or 1, according to the absence or presence of dysphonia, respectively. The absence or presence of the vocal disorder was determined by the outcome of the videolaringostroboscopy examination. Hence, the final step of the first part of this work is the statistical analysis, which aims to look for the best indicator of vocal health status among the acoustic parameters. Actually, this study takes into account also classification models based on couples of parameters, which could lead to enhanced results.

Several statistical tests have been conducted through the software RStudio®. A Generalized Linear Model has been performed; it is a logistic regression model, applied when the response variable (in this case the voice health status) is not normally distributed. When the response variable, that is the *dependent variable*, is defined by a dichotomous outcome, 0 or 1, the logistic regression is appropriate. The function *glm* of RStudio allows creating different models, i.e. different potential classifiers. In particular, it is possible to perform a *single-variable* logistic regression model for each parameter, or a *multiple* one. The former allows for the model building by a single *independent variable*; on the other hand, the multiple logistic regression model considers more than one independent variables. Anyway, the independent variables have to be chosen from the set of the available features. Particularly, for a multiple model it is necessary pay attention on the correlation between the parameters used in the model construction, since they could be carriers of the same information, without improving the outcomes. For example, in this study, a single-variable logistic regression model has been performed for each feature previously extracted, but also a two-variable one, choosing among the possible couples of parameters those that showed a low correlation. The intent would be to combine information from two different parameters in order to achieve a model more accurate than a single-variable one.

However, through one or more independent variables and involving the whole database (training set), a function of probability is modelled (with a link function *logit*). It provides the probability of an output variable to be equal to one, that is in this case the probability for a subject to be unhealthy. For each regression model tested, the algorithm returns the intercept and the slope relative to each independent variable, thanks to which it is possible to define the function of probability as:

$$P(Unhealthy = 1) = \frac{e^{(Intercept + Slope * Parameter)}}{1 + e^{(Intercept + Slope * Parameter)}}$$

$$P(Unhealthy = 1) = \frac{e^{(Intercept + Slope1 * Parameter1 + Slope2 * Parameter2)}}{1 + e^{(Intercept + Slope1 * Parameter1 + Slope2 * Parameter2)}}$$

Where the first expression is referred to a single-variable logistic regression model, while the second one to a two-variable logistic regression model. *P(Unhealthy)* is the probability of having unhealthy voice and ranges from zero to one.

After that, by means of some functions, the statistical software permits the performance evaluation of the model. In this way, it is possible to compare different models and select the best one.

Firstly, through the command *summary*, the Akaike information criterion (AIC) is returned. The AIC is an estimator of the relative quality of a specific dataset. It not gives absolute information about the model, but it estimates the quality of a parameter (or of a couple of parameters) in relation to the other parameters. It keeps in mind the lost information when a model is used, compared to all the complexity of the model. Once obtained the AIC values for all the considered models, the best one is that with minor value of AIC. Also the value R^2 of McFadden can be calculated. It gives a goodness estimation of the logistic regression model. The value is much closer to 1, when the model is best. In order to assess the differentiation between two subject groups, a further check has been considered, but only for single-variable models: the test Wilcoxon (or U-test).

This test is not parametric for dependent samples. The null hypothesis of the test establishes that the two groups (healthy and unhealthy) belong to the same population and consequently their probability distribution is the same. The returned p-values are compared with the value of 0,05. If the p-value related to a parameter is lower than the threshold value, the null hypothesis is rejected and the probability distribution of that parameter can be considered enough different for the two groups. Therefore, the results of U-test can help in the choice of the best parameter, evaluating the lower ones.

When the study object is a diagnostic test based on dichotomous outcome (positive or negative, healthy or unhealthy), like in this case-study, there is another approach to evaluate the model predictive power, which is the most widely used in literature: it is linked to the concept of *sensitivity* and *specificity* [40]. The sensitivity is the true positive rate, i.e. probability to identify correctly the pathological subjects, while the specificity is the true negative rate, i.e. the probability to identify correctly the subjects without voice problems. Their values depend on the threshold (*cut-off*) that determine the positive-negative classification. For the logistic regression model, the threshold value ranges between 0 and 1. Once calculated intercept and slope of the model, a value of *P(Unhealthy)* is assigned to all the subjects of the training set: a *P(Unhealthy)* value higher than the threshold means pathological voice (positive), on the contrary, a lower one means healthy voice (negative). After that, sensitivity and specificity are evaluated as the threshold changes. The relationship between sensitivity and 1-specificity (i.e. false positive rate) is described by the ROC Curve (fig.3.8). Therefore, the ROC curve is realized from the probability distribution of healthy and unhealthy group. A perfect discriminating test is represented by a complete separation of the two distributions, on the contrary, a complete overlap cannot allow to use the test. Step by step, the cut-off is defined and the corresponding sensitivity and specificity become the coordinates of the curve itself. The Area Under Curve (AUC) is representative of the diagnostic test performance, because it is related to the location of the ROC Curve. It ranges from 0,5 to 1 and describes the classification accuracy of the model, so it can be

indicated by also percentage. An AUC near to 1 suggests a strong model ability to separate subjects with vocal disorders from the ones with normal voices; differently, an AUC close to 0,5 identifies low capability to distinguish between the two groups, because the probability to classify a subject 0 or 1 is the same. In the choice of the best model, high value of AUC is decisive.

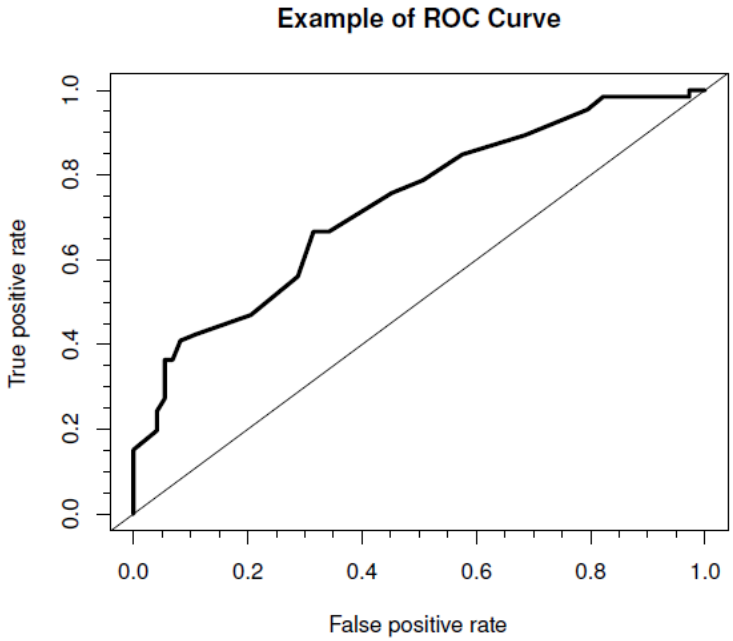


Figure 3.8: Example of ROC Curve; True positive rate is Sensitivity and False positive rate is 1-Specivicity.

According to several statistical tests (for example []), the AUC result can be interpreted as follows:

AUC = 0.5 failed test;

$0.5 < AUC \leq 0.7$ little accurate test;

$0.7 < AUC \leq 0.9$ moderate accurate test;

$0.9 < AUC < 1$ highly accurate test;

AUC = 1 perfect test.

Conventionally, a diagnostic test can be considered relevant for $AUC \geq 0.80$ [41].

Overall, the evaluation of the statistical results (AIC, McFadden, U-test, AUC) allows to compare each other all the examined models and select the parameter or the couple of parameters as the best predictor of voice health status.

3.2.6 Cut-off evaluation and model validation

Once compared each other all the tested models and identified, where possible, the one with the best predictive power for a certain voice task and a certain microphone, the analysis carries on with the cut-off selection only for the selected model. This is the final step leading to the real construction of a healthy-unhealthy classifier. Usually, the cut-off is detected by the intersection between sensitivity and specificity curves (fig.3.9), but, according to the authors, priority should be given to sensitivity. As said earlier, the detected cut-off corresponds to a probability. Therefore, in a two-variable model it remains so; by contrast, for a single-variable model, it is possible to express the threshold in terms of the parameter (with the same unit of measure), using the first formula reported in the previous paragraph.

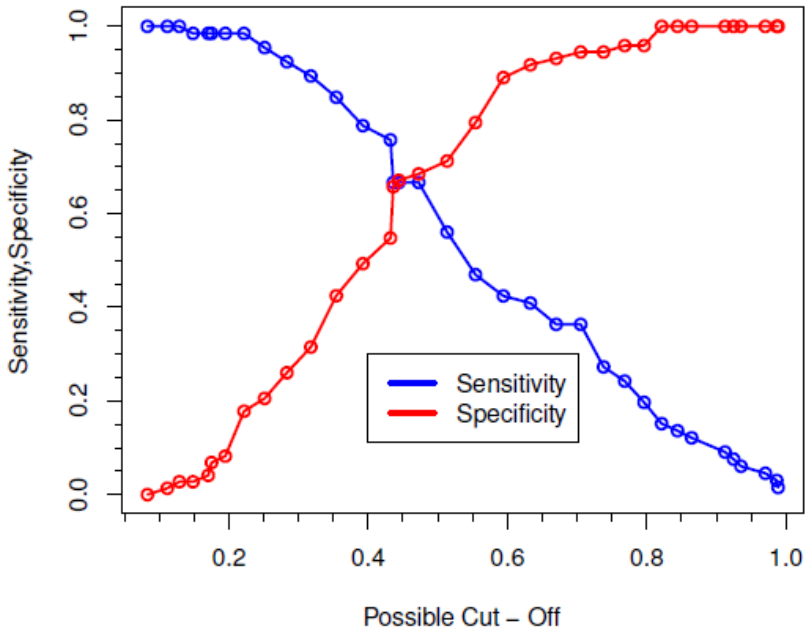


Figure 3.9: Sensitivity and Specificity curves versus possible cut-off value.

This phase coincides with the classifier validation, which consists in the assessment of its performance, mainly in terms of accuracy. The accuracy can be defined as the percentage of cases correctly classified, independently from healthy and unhealthy subjects. Selecting the cut-off, at the same time, the sensitivity and specificity values of the model are found. As suggested in the previous section, they are accuracy measures: in fact, sensitivity is the percentage of cases correctly classified as positive, while specificity to the percentage of cases correctly classified as negative. Eventually, from their values, it is possible to calculate the real accuracy of the model. This just explained, is a validation that consists in the classification of subjects from the dataset used to find the model. In detail, the threshold selection determines their positive-negative classification, so the performance of the classifier.

In this study another type of validation has been also performed. As said in paragraph, a preliminary analysis has considered the whole dataset as divided into two groups: training set and test set. The former, includes the data used in the model fitting; the latter, includes data that has not been involved in the building of the classifier. The subjects of the test set, in a following step, have been classified using the results obtained with the training set. The two dataset are so composed:

- Training set: 77 patients and 64 controls, overall 141 subject (80% of the whole dataset).
- Test set: 25 patients and 9 controls, overall 34 subjects (20% of the whole dataset).

The reason that explains this distinction lies in a previous work, where some models had been selected as the best ones in differentiating normal voices from dysphonic voices, in both sustained vowels and continuous speech and for the different microphones. The previous study had been conducted using only the 141 subjects, as the training set. Only single-variable regression models had been performed and, as regards sustained vowels, only CPPS parameters had been investigated. The cut-off detection and the performance evaluation had been

executed using the same dataset (test set = training set). In detail, four models had provided satisfying outcomes for:

1. Sustained vowel /a/ with microphone in air;
2. Sustained vowel /a/ with ECM microphone;
3. Reading with microphone in air;
4. Free speech with microphone in air.

Therefore, the preliminary analysis started from these existing models applying them to the 34 subjects that in this case represent the test set, in order to evaluate their *generalization ability*. In fact, since each final model had been defined through the classification of the subjects from the training set, now its performance could strictly depend on the data used for its creation. For this reason, another validation consists in using the results previously achieved to classify new subjects. By means of a *Confusion Matrix* (CM) related to the test set classification, the generalization ability of the existing models have been evaluated in terms of accuracy, percentage of cases correctly classified as positive and percentage of cases correctly classified as negative: these values have to be compared with the ones found from the training set classification.

The next step was the research of new models with better performance. For the four cases above mentioned, the separation between training set and test set was kept and several two-variables logistic regression models were tested. Moreover, only for sustained vowels, the perturbation parameters and HNR were added in the investigation of the best predictor. By contrast, for Reading and Free speech acquired with ECM microphone and for all the different voice tasks of the piezoelectric microphone, the whole dataset was used as training set and test set, since good models had not been found in the previous work. Also in this case two-variable models and additional parameters were considered.

3.3 Sustained vowel analysis

This chapter concerns the results and discussion about the analysis of the vowel /a/ signals. For each audio file, the nine descriptive statistics of the CPPS distribution, the nine perturbation measures and the HNR parameter have been calculated and collected in a database for the following statistical analysis, which aims to search for models able to distinguish healthy from unhealthy voices. In order to simplify the investigation, a preliminary feature selection has been performed on the perturbation parameters and HNR, by evaluating the Pearson correlation coefficient between the parameters, considered two at a time. Then, the existing models found in a previous study for sustained vowel, have been validated testing their generalization ability. Furthermore, other models both single- and two-variable have been proposed for the dysphonia recognition. These analyses have been conducted for all the three different microphones.

3.3.1 Feature selection

***Pearson's correlation coefficient (ρ)* between vowel /a/ parameters**

vowMIPRO vowPIEZO	Jitt	Jita	RAP	PPQ	vF0	Shim	ShdB	APQ	vAm	HNR	CPPS 5prc	CPPS 95prc	CPPS Std
Jitt		0,99 0,99	0,99 0,99	0,69 0,52	0,89 0,92	0,89 0,74	0,88 0,72	0,88 0,73	0,42 0,25	-0,54 -0,64	-0,67 -0,52	-0,54 -0,57	0,67 0,46
Jita			0,99 0,99	0,71 0,52	0,89 0,91	0,89 0,78	0,89 0,78	0,9 0,77	0,42 0,25	-0,66 -0,63	-0,74 -0,54	-0,55 -0,43	0,66 0,48
RAP				0,69 0,52	0,89 0,92	0,89 0,73	0,88 0,71	0,89 0,72	0,42 0,24	-0,63 -0,59	-0,73 -0,5	-0,54 -0,4	0,66 0,45
PPQ					0,54 0,49	0,68 0,36	0,48 0,32	0,66 0,36	0,15 0,07	-0,47 -0,33	-0,53 -0,25	-0,56 -0,29	0,66 0,12
vF0						0,83 0,71	0,85 0,72	0,84 0,72	0,51 0,29	-0,69 -0,66	-0,76 -0,5	-0,49 -0,35	0,75 0,49
Shim							0,99 0,99	0,99 0,99	0,49 0,49	-0,81 -0,8	-0,76 -0,77	-0,63 -0,58	0,61 0,7
ShdB								0,98 0,99	0,49 0,55	-0,81 -0,81	-0,76 -0,78	-0,63 -0,58	0,67 0,72
APQ									0,49 0,5	-0,82 -0,82	-0,77 -0,79	-0,63 -0,72	0,63 0,72
vAm										-0,48 -0,32	-0,56 -0,55	-0,28 -0,55	0,64 0,35
HNR											0,74 0,72	0,55 0,45	-0,63 -0,77

Figure 3.9: Feature selection.

Subset of parameters selected: Jitt, PPQ, Shim, vAm and HNR.

3.3.2 Microphone in air: validation of existing model

For this analysis, the whole dataset was distinguished between two groups: the training set (80% of the whole dataset) and the test set (20% of the whole dataset). The former had been just used in a previous work and table 3.4 summarizes the results of the statistical analysis for sustained vowel acquired with the microphone in air (MIPRO). Only the CPPS parameters had been investigated. The criteria that have been followed by both the antecedent and actual study for the choice of the best feature are:

Low AIC

High coefficient of McFadden R^2

p-value of U-test < 0,05

high AUC

CPPS parameter	AIC	Mcfadden R^2	U-test	AUC
CPPS mean	95,3	0,42	0,000	0,90
CPPS median	98,9	0,39	0,000	0,89
CPPS mode	104,0	0,36	0,000	0,87
CPPS std	127,7	0,21	0,000	0,80
CPPS range	137,9	0,15	0,000	0,74
CPPS 5prc	**88,6**	**0,46**	**0,000**	**0,91**
CPPS 95prc	105,7	0,35	0,000	0,87
CPPS skewness	159,9	0,01	0,096	0,59
CPPS kurtosis	160,2	0,01	0,599	0,47

Table 3.4: Results of the previous statistical analysis for sustained vowel in MIPRO.

The most important criterion is the one relative to the Area Under Curve (AUC), whose value can considered as a real index of the model predictive power. Taking this into account, from the previous statistical results, the 5^{th} percentile of the CPPS distribution ($CPPS_{5prc}$) had proved to be the best parameter in discriminating between healthy and pathological voices. Table 3.5 illustrates the

characteristics of this model: intercept, slope, cut-off (in terms of P(Unhealthy) and in dB), Confidence Interval (CI, in dB), sensitivity and specificity. In particular, sensitivity and specificity, which are performance indicators, were equal to 0.85 and 0.81, respectively. Knowing the number of true healthy and true unhealthy subjects, the model accuracy can be also estimated; it was equal to 83%. The confidence interval results from an uncertainty estimation of the threshold value, performed in the antecedent study by means of Monte Carlo method.

Best model	Int.	Slope	Th (P(U))	Th(dB)	CI (dB)	Sens.	Spec.	Acc.
CPPS 5prc	13,7	-0,96	0,6	13,8	1,06	0,85	0,81	83%

Table 3.5: Characteristics and performance of the existing model for sustained vowel acquired with MIPRO microphone.

The following expression defines the best empirical fitted model that was found:

$$P(Unhealthy = 1) = \frac{e^{(13{,}7-0{,}96*CPPS_{5prc})}}{1 + e^{(13{,}7-0{,}96*CPPS_{5prc})}}$$

where *P(Unhealthy)* is the probability of having unhealthy voice and ranges between 0 and 1, as explained in paragraph 3.2.5.

The next step is the validation of the classifier using the new dataset of 34 subjects, called test set. The equation above reported was applied to all the samples of the test set, assigning to each one a probability to be unhealthy. Thanks to the cut-off value previously extracted, a classification of the 34 subjects was made. Then, knowing their real voice health status, a Confusion Matrix was computed (fig. 3.10), from which it was possible to extract the accuracy, sensitivity and specificity of the new classification. They resulted equal to about 82%, 80% and 89%, respectively. The aim was to evaluate with which performance the found model is able to assign a class to subjects that had not been involved in the training phase. Since all the three values are not so different from

the ones obtained classifying the training set, it is possible to assert that the existing model has proved a good generalization capability.

CM	Unealthy (25)	Healthy (9)
Positive	20	1
Negative	5	8

$$Accuracy = \frac{20+8}{25+9} * 100 = 82{,}3\%$$

$$Sensitivity = \frac{20}{25} * 100 = 80\%$$

$$Specificity = \frac{8}{9} * 100 = 88{,}9\%$$

Figure 3.10: Confusion Matrix from the test set classification.

Fig. 3.11 shows a graph where the probability to be unhealthy is reported for each subject of the test set. As expected, the healthy subjects (green crosses) are grouped below the threshold (red line), while pathological ones over the threshold, and some of the latter have precise value 1. Moreover, for unhealthy subjects, the qualitative parameter G of GIRBAS scale has been considered: most not correctly classified subjects have G equal to 1, all subjects with G equal to 2 have been correctly classified, while among the subjects with G equal to 3, one has not been correctly classified. So, the classifier has not showed ability in generalizing about the most serious patients. The meaning of G discrimination in this type of graph is that any model should tend to assign the wrong class to G1 unhealthy subjects, because they have a mild grade of dysphonia and very often they compensate the disease, while should not make many mistakes in classifying G2 and G3 patients. In the graph also the confidence interval (dotted red lines) is showed for a complete representation: subjects included within this interval are not clearly classified.

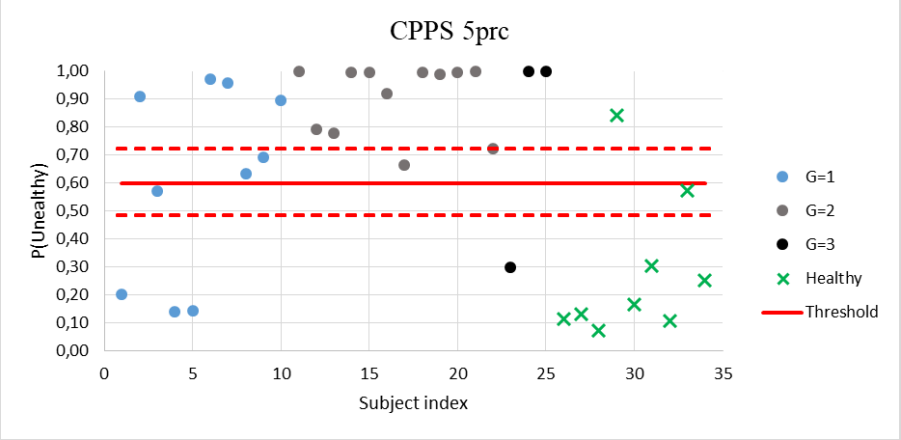


Figure 3.11: Test set classification in terms of P(Unhealthy) in comparison with G value.

3.3.3 Microphone in air: a new model

The next step dealt with a further investigation on looking for, if possible, a new model with greater performance than the earlier one. For this purpose, also the perturbation parameters and the HNR were involved in the analysis. Obviously, only the ones identified by the feature selection were considered: Jitt, PPQ, Shim, vAm and HNR. Furthermore, both single- and two-variable logistic regression models were experimented, with the idea, in the second case, of combining the information from two not correlated parameters and improving the prediction power.

In this phase, the separation of the data set in training and test set was maintained. As a consequence, the statistical analysis through a single independent variable was performed only on the additional parameters, while the one through two independent variables involved also some of the nine CPPS statistics.

Table 3.6 summarizes the results relative to the effectiveness of the models obtained assuming the presence-absence of vocal disorders as dependent variable and the parameters, one at a time, as independent variables. According to the values of the statistical indicators, the PPQ parameter has satisfied the criteria more than the others. In particular, it provided an AUC of 0.84, which is over the AUC acceptable value in a diagnostic test (0.80). However, the outcomes of PPQ did not overcome the ones returned by the $CPPS_{5prc}$ in the earlier investigation.

Parameter	AIC	Mcfadden R^2	U-test	AUC
Jitt	144,3	0,07	0,000	0,78
PPQ	127,4	0,18	0,000	0,84
Shim	142,7	0,08	0,000	0,83
vAm	139,2	0,11	0,000	0,79
HNR	127,0	0,19	0,000	0,76

Table 3.6: Results of statistical analysis using perturbation parameters and HNR.

The analysis continued with testing two-variable regression models, choosing couples of little correlated parameters. In detail, the following strategy was adopted: among the CPPS parameters that have produced good results, mean, median, mode, 5^{th} percentile and 95^{th} percentile stand out, but they are all location measures of the CPPS distribution, so it is supposed that they are highly correlated. As a result, only the $CPPS_{5prc}$, which was the best, has been compared with each new parameter through the Pearson's correlation coefficient ρ; the outcomes are showed in tab. 3.7 and highlight the great correlation of $CPPS_{5prc}$ with HNR (red) and the poor correlation with the other four parameters. Therefore, the following couples have been selected for the statistical analysis: $CPPS_{5prc}$+Jitt, $CPPS_{5prc}$+PPQ, $CPPS_{5prc}$+Shim, $CPPS_{5prc}$+vAm. The values of statistical indexes for each two-variable model are reported in tab. 3.8. It is evident that all the models reach satisfying results; in particular, the AUC value is over 0.90 for all ones. However, the 5^{th} percentile of the CPPS distribution combined with the PPQ perturbation parameter exhibits the best outcomes for all the statistical criteria considered; this result find a correspondence in the lowest correlation showed by $CPPS_{5prc}$ and PPQ (tab 3.7, green). Furthermore, this two-variable model has an AUC equal to 0.92 (highly accurate test), greater than the AUC of the model obtained by the single $CPPS_{5prc}$. The same consideration is valid for also the other statistical indicators: it can be assumed as the best predictor of vocal health status.

Correlation (ρ)	Jitt	PPQ	Shim	vAm	HNR
CPPS 5prc	**-0,60**	**-0,21**	**-0,55**	**-0,43**	**0,82**

Table 3.7:

Two parameters	AIC	Mcfadden R^2	AUC
CPPS 5prc + Jitt	88,0	0,46	0,91
CPPS 5prc + PPQ	**83,2**	**0,49**	**0,92**
CPPS 5prc + Shim	87,8	0,46	0,91
CPPS 5prc + vAm	87,3	0,46	0,91

Table 3.8:

The formula that defines the best empirical fitted logistic model is:

$$P(Unhealthy = 1) = \frac{e^{(12{,}4-0{,}89*CPPS_{5prc}+0{,}08*PPQ)}}{1 + e^{(12{,}4-0{,}89*CPPS_{5prc}+0{,}08*PPQ)}}$$

The statistical analysis conducted with RStudio has provided for this model the ROC Curve depicted in fig. 3.12. The next step is the evaluation of the threshold in order to accomplish the positive-negative classifier: from the probability distribution of healthy and unhealthy subjects, the cut-off has been varied and for each value, the relative sensitivity and specificity have been calculated. Fig. 3.13 consists in the graph that gives information about sensitivity and specificity in function of the possible cut-off value. The threshold has been searched in the graph where sensitivity and specificity are similar, privileging a greater sensitivity. In the specific case, the choice fell on a threshold value of 0.54, expressed in terms of probability. At this value, sensitivity and specificity are approximately 0.85 and 0.81, leading to an accuracy equal to 83%. The overall characteristics and performance of the selected model are reported in tab. 3.9.

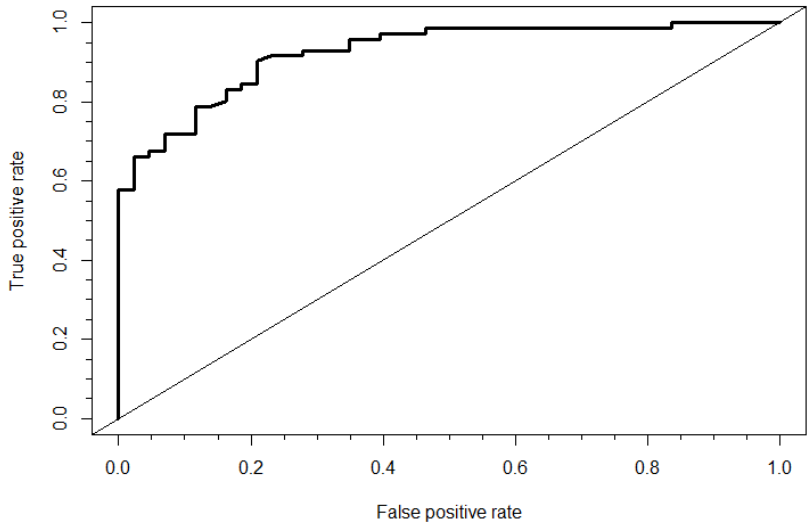


Figure 3.12:

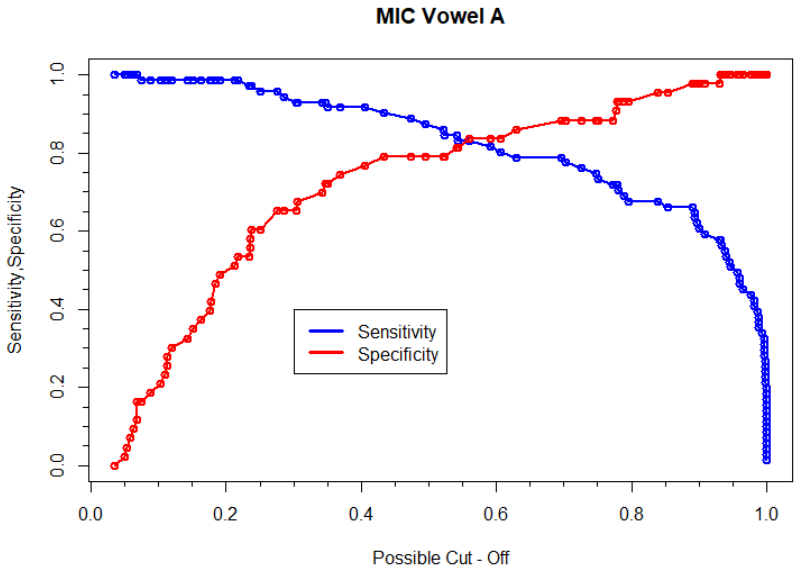


Figure 3.13:

Best model	Int.	Slope 1	Slope 2	Th (P(U))	Sens.	Spec.	Acc.
CPPS 5prc + PPQ	12,4	-0,89	0,08	0,54	0,85	0,81	83%

Table 3.9:

Fig. 3.14 shows the graph in which for each subject from the training set the probability to be pathological one is represented: as expected, most patients are in the top part, where the probability is near to 1, while most controls are in the bottom part, near to 0. An important evidence is that all the unhealthy subjects who are wrongly classified by the model correspond to those who were judged with the lowest overall grade G of dysphonia.

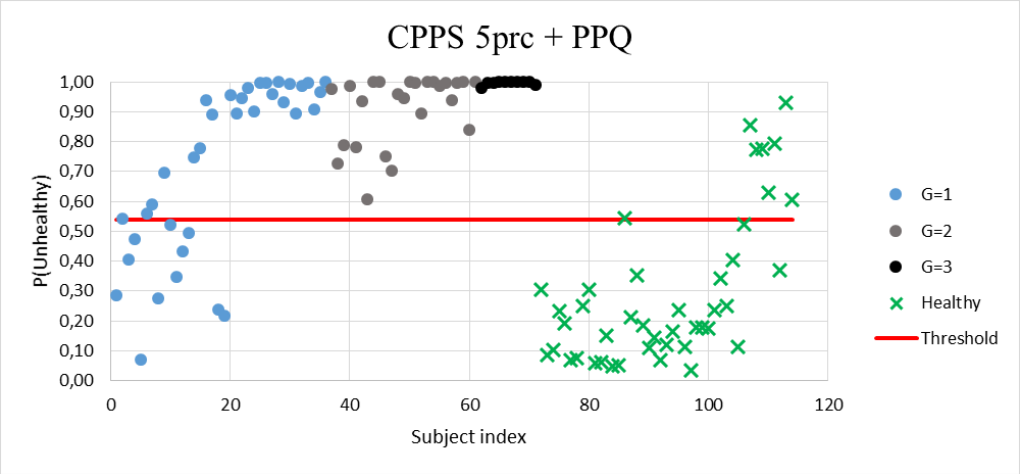


Figure 3.14:

It is evident that evaluating the model performance by classifying the subjects from the training set, better results are not obtained, in comparison with the performance of the earlier single-variable model. In fact, both classifiers are characterized by the same sensitivity, specificity and accuracy. However, the following phase is the validation of the new model from generalization ability point of view. Applying the ($CPPS_{5prc}$+PPQ)-model to the subjects of the test set using the formula above expressed, and classifying them according to the selected threshold, the Confusion Matrix represented in fig. 3.15 is obtained. The following considerations can be extracted (in percentage terms): from the comparison with the $CPPS_{5prc}$-model, it emerges a better capability in classifying correctly pathological voices (sensitivity passes from 80% to 84%), while specificity has maintained the same value. Overall, accuracy, i.e. the percentage of correctly cases classified as positive and negative, has increased from 82% to 85%, demonstrating that the generalization ability has improved with the new

two-variable model. Finally, the graph that displays the probability of having unhealthy voice for each subject of the test set is provided (fig. 3.16): in particular, by comparison it with the graph in fig. 3.11, a relevant result is that the subject with G equal to 3 (severe overall grade of dysphonia), now is correctly identified as positive. In addition, the two healthy-unhealthy groups are well separated.

CM	Unhealthy (25)	Healthy (9)
Positive	21	1
Negative	4	8

$$Accuracy = \frac{21+8}{25+9} * 100 = 85{,}3\%$$

$$Sensitivity = \frac{20}{25} * 100 = 84\%$$

$$Specificity = \frac{8}{9} * 100 = 88{,}9\%$$

Figure 3.15:

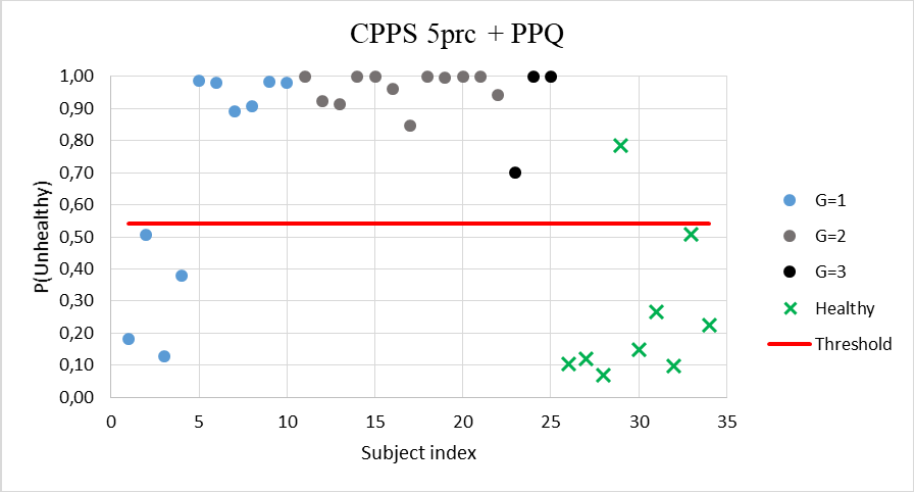


Figure 3.16:

3.3.4 Microphone ECM: validation of existing model

Table 3.10 summarizes the statistical results which had been reached with the earlier model: the $CPPS_{5prc}$ had been found as the best parameter in discriminating healthy voices from dysphonic ones, in sustained vowel acquired with the contact microphone ECM. The AUC value of the selected model was 0.82, which indicates a moderate discrimination power. In table 3.11 the characteristics and performance of the model are summarized.

CPPS parameter	AIC	Mfadden R^2	U-test	AUC
CPPS mean	131,1	0,22	0,000	0,78
CPPS median	133,0	0,20	0,000	0,78
CPPS mode	136,7	0,18	0,000	0,75
CPPS std	130,2	0,22	0,000	0,80
CPPS range	143,8	0,14	0,000	0,77
CPPS 5prc	**121,6**	**0,27**	**0,000**	**0,82**
CPPS 95prc	143,4	0,14	0,000	0,72
CPPS skewness	164,5	0,01	0,778	0,49
CPPS kurtosis	162,1	0,02	0,915	0,50

Table 3.10

Best model	Int.	Slope	Th (P(U))	Th(dB)	CI (dB)	Sens.	Spec.	Acc.
CPPS 5prc	7,5	-0,47	0,46	16,3	1,29	0,73	0,73	73%

Table 3.11:

Also in this case, in order to test the generalization ability of the selected model, the formula that allows calculating the fitted values of subjects in terms of probability to be pathological, combined with the selected cut-off, was used to classify the subjects of the test set. The performance evaluation of the classifier is possible through the Confusion Matrix depicted in fig. 3.17: the values of

sensitivity, specificity and accuracy are lower than the ones obtained from the training set classification, indicating poor generalization capability of the model. Such a result is also evident by observing the graph in fig. 3.18, where the two groups of healthy and unhealthy subjects are not well detached; in fact, most of them have a P(Unhealthy) close to the threshold value.

CM	Unhealthy (25)	Healthy (2)
Positive	16	1
Negative	9	1

$$Accuracy = \frac{16+1}{25+2} * 100 = 63\%$$

$$Sensitivity = \frac{16}{25} * 100 = 64\%$$

$$Specificity = \frac{1}{2} * 100 = 50\%$$

Figure 3.17:

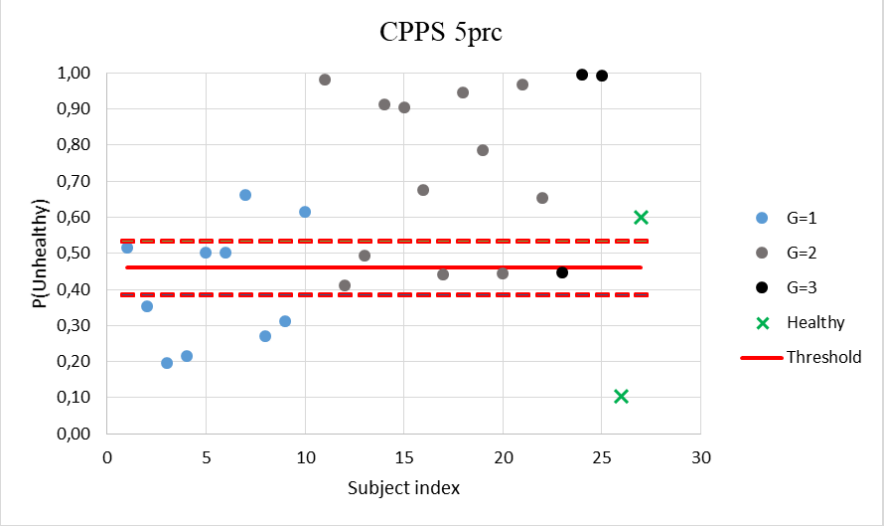


Figure 3.18:

3.3.5 Microphone ECM: a new model

After the performance evaluation of the antecedent model, on the same dataset, five single-variable logistic regression models were implemented through the parameters Jitt, PPQ, Shim, vAm and HNR. The values of the statistical indicators are reported in tab. 3.12: it is possible to observe that PPQ is the parameter with greater value of R^2 (0.22) and AUC (0.80).

Parameter	AIC	Mcfadden R^2	U-test	AUC
Jitt	149,8	0,09	0,000	0,75
PPQ	128,8	0,22	0,000	0,80
Shim	153,0	0,07	0,000	0,74
vAm	146,5	0,11	0,000	0,73
HNR	153,1	0,07	0,019	0,63

Table 3.12:

The investigation carried on with the two-variable approach, always considering the same training set. Table 3.13 shows the Pearson coefficients calculated to quantify the correlation between $CPPS_{5prc}$ and the five additional parameters, and between $CPPS_{std}$ and the same. Among the CPPS statistics, also standard deviation was considered, since it was the second best discriminant parameter (tab. 3.10). For the couples that proved to be little correlated, the models with two independent variables have been experimented; table 3.14 summarizes the relative results. For all the tested pairs, the AUC is higher than or equal to 0.80, and also the values of Mc Fadden's R^2 are relatively great indicating a good separations between patients and controls: these results are consistent with the idea of combining two parameters that convey different information of the vocal signal. However, the couple composed by $CPPS_{std}$ and PPQ has the best outcomes (R^2 equal to 0.34 and AUC equal to 0.85), the same if compared to the earlier model; it exhibites a moderate discrimination power. Such a result find a justification in the lowest Pearson coefficient between CPPS standard deviation and PPQ parameter (tab. 3.13, green).

Correlation (ρ)	Jitt	PPQ	Shim	vAm	HNR
CPPS 5prc	-0,66	-0,61	-0,61	-0,65	0,73
CPPS std	0,36	0,28	0,41	0,61	-0,4

Table 3.13:

Two parameters	AIC	Mcfadden R^2	AUC
CPPS 5prc + Jitt	123,1	0,27	0,82
CPPS 5prc + PPQ	115,3	0,32	0,83
CPPS 5prc + Shim	123,1	0,27	0,82
CPPS 5prc + vAm	122,7	0,27	0,82
CPPS std + Jitt	127,9	0,24	0,82
CPPS std + PPQ	**111,9**	**0,34**	**0,85**
CPPS std + Shim	128,3	0,24	0,80
CPPS std + vAm	126,5	0,25	0,81
CPPS std + HNR	129,6	0,23	0,80

Table 3.14:

The next step for the building of a new classifier consists in calculating the fit values for each subject of the training set, by means of the following formula:

$$P(Unhealthy = 1) = \frac{e^{(-3,6+2,35*CPPS_{std}+0,12*PPQ)}}{1 + e^{(-3,6+2,35*CPPS_{std}+0,12*PPQ)}}$$

that highlights the values of intercept and slopes of the empirical model.

After that, the threshold evaluation has been conducted, by observing the graph where sensitivity and specificity versus each possible cut-off value are plotted (fig. 3.19). The best classification threshold is *P(Unhealthy)* = 0.37, with a sensitivity of 0.85, a specificity of 0.69 and an accuracy of 77%. Table 3.15 reports the characteristics and performance of the new model. Sensitivity is greater than the one obtained with the previous model (0.73), considering the training set classification; by contrast, specificity is lower. All in all, accuracy, in percentage terms, is improved from 73% to 77%. Fig. 3.19 also shows that most

patients that are wrongly classified by the model have been perceptually rated with the lowest overall grade G of dysphonia.

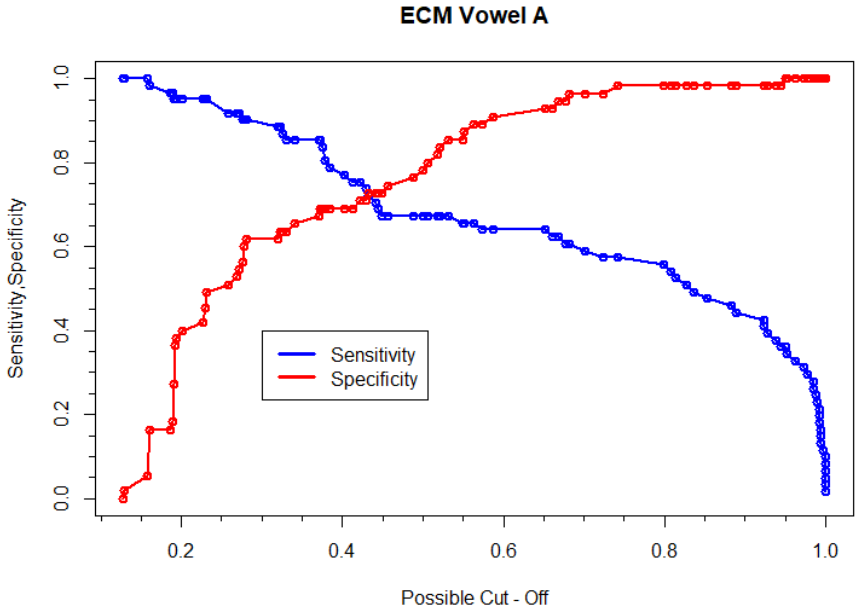


Figure 3.19:

Best model	Int.	Slope 1	Slope 2	Th (P(U))	Sens.	Spec.	Acc.
CPPS std + PPQ	-3,6	2,35	0,12	0,37	0,85	0,69	77%

Table 3.15

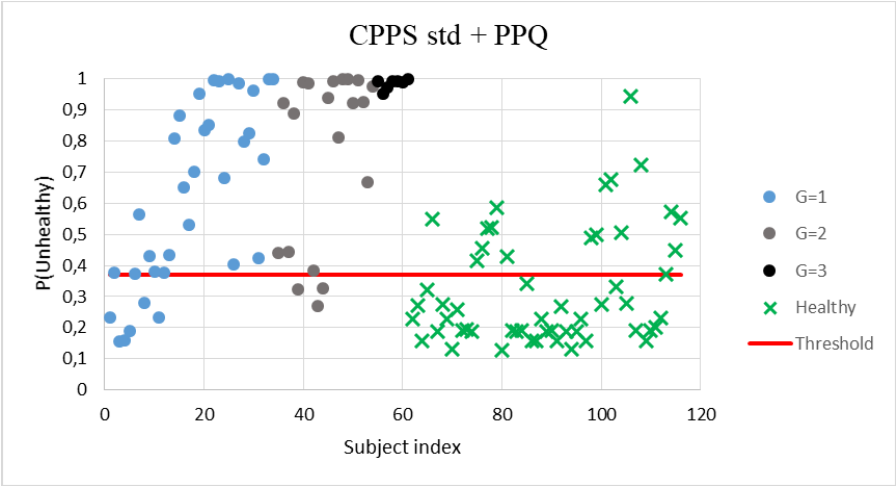


Figure 3.20:

The final step was the evaluation of the model generalization ability. The subjects of the test set have been classified using the outputs of the model (intercept, slopes relative to the two independent variables) and the cut-off previously selected. The Confusion Matrix was calculated for the model validation and the values of sensitivity, specificity and accuracy have been extracted (fig. 3.21, in percentage terms) and compared with those from the training set classification. The accuracy has rose dramatically from 77% to around 89%. Since specificity has not changed, this effect is linked to a considerable increase in sensitivity (from 85% to 92%). Therefore, thanks to the new two-variable model, the $CPPS_{std}$-PPQ model, an outstanding improvement in the classification of subjects not involved in the training step has occurred. These considerations are also clear from the graph in fig. 3.22: only two unhealthy subjects, with G equal to 1, have been wrongly classified.

CM	Unhealthy (25)	Healthy (2)
Positive	23	1
Negative	2	1

$$Accuracy = \frac{23+1}{25+2} * 100 = 88{,}9\%$$

$$Sensitivity = \frac{23}{25} * 100 = 92\%$$

$$Specificity = \frac{1}{2} * 100 = 50\%$$

Figure 3.21:

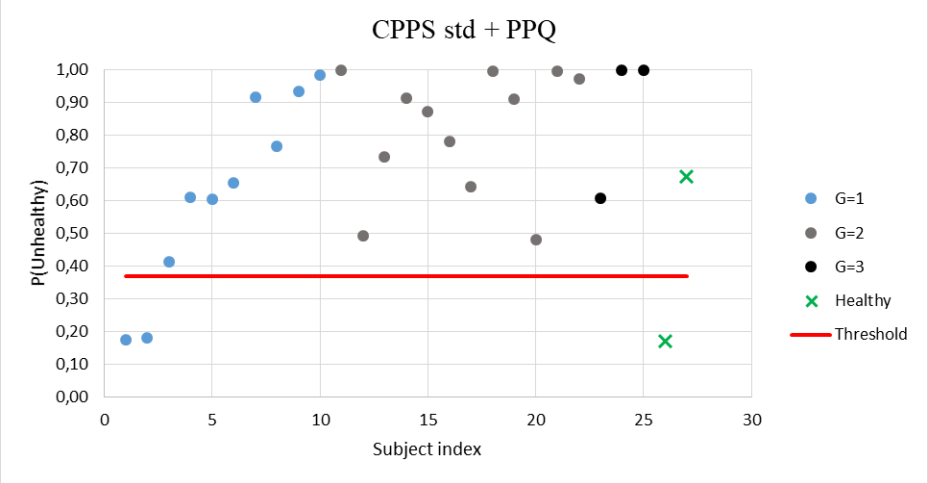


Figure 3.22:

3.3.6 Piezoelectric microphone: results

As regards the analysis of sustained vowel /a/ recorded with the piezoelectric microphone, preliminary studies had not led to a final classifier, since none of the nine CPPS statistics had showed relevant results in discriminating between normal voices and dysphonic voices. The reason lies in the fact that the dataset related to piezoelectric microphone and used in the earlier work was composed by about 52 subjects (26 patients and 26 controls). The number is less consistent for statistical analysis and the results cannot be compared with the other microphones, which had a database composed at least by twice the amount of subjects. However, it seemed that the outcomes, in terms of values of the statistical indexes, tended to confirm the ones obtained by the microphone in air and from the other contact microphone ECM: the $CPPS_{5prc}$ exhibited results slightly better than other CPPS parameters.

The subdivision of the dataset into two groups, training set and test set, is not valid in this case. For the healthy-unhealthy discrimination of piezoelectric microphone in sustained vowel (but also in continuous speech), in this work, the whole available dataset has been used for the statistical analysis: 102 patients and 73 controls. Tab. 3.16 illustrates the results for single-variable logistic regression models related to the nine CPPS descriptive statistics and the five parameters that have been added in this work, only for sustained vowel. It is possible to observe that, among the CPPS statistics, the $CPPS_{5prc}$ is again the one that comes closest to the statistical criteria. However, among the five additional parameters, there is the PPQ that can be considered the best discriminant parameter, with a McFadden R^2 equal to 0.37 and an AUC of 0.88, which suggest a moderate accuracy in the separation between patients and controls.

Parameter	AIC	Mcfadden R^2	U-test	AUC
CPPS mean	100,0	0,16	0,000	0,74
CPPS median	100,4	0,16	0,000	0,74
CPPS mode	101,0	0,15	0,000	0,74
CPPS std	99,2	0,17	0,000	0,77
CPPS range	103,4	0,13	0,000	0,74

CPPS 5prc	92,0	0,23	0,000	0,79
CPPS 95prc	108,1	0,09	0,010	0,67
CPPS skewness	115,9	0,02	0,800	0,51
CPPS kurtosis	115,5	0,03	1,000	0,50
Jitt	113,1	0,05	0,001	0,72
PPQ	76,5	0,37	0,000	0,88
Shim	116,6	0,02	0,001	0,71
vAm	114,4	0,04	0,000	0,75
HNR	111,9	0,06	0,060	0,62

Table 3.16:

The next step consisted in calculating the Pearson correlation coefficients between the $CPPS_{5prc}$ and the other parameters Jitt, PPQ, Shim, vAm and HNR (table 3.17), in order to select the couples of parameters through which two-variables models have been experimented. Table 3.18 indicates what combinations have been considered (only the $CPPS_{5prc}$-HNR couple has not be tested due to the high correlation, red in table) and what the corresponding results have been had. The couple made up of the 5^{th} percentile of CPPS distribution and the PPQ perturbation parameter has produced the best logistic regression model as indicator of dysphonia severity. Such a result has not found a conformity in the Pearson coefficient between $CPPS_{5prc}$ and PPQ, since they were not the less correlated parameters (table 3.17, green); probably the already high discrimination power of PPQ overcame this limit.

The analysis could carry on with the cut-off selection, by means of the graph reported in fig 3.23.: the choice fell on *P(Unhealthy)* = 0.5, with a sensitivity of 0.84 and a specificity of 0.77. The accuracy, in percentage terms, was 81%. The overall characteristics of the model are described in table 3.19. In particular, it is possible to notice that the threshold (in terms of probability of having unhealthy voice) is lower than the one found for the microphone in air. This is because the latter microphone is not affected by external noise, which influences the CPPS distribution with lower values, leading to high values of *P(Unhealthy).*

	Jitt	PPQ	Shim	vAm	HNR
CPPS 5prc	**-0,66**	**-0,67**	**-0,50**	**-0,54**	**0,80**

Table 3.17:

Two parameters	AIC	Mcfadden R^2	AUC
CPPS 5prc + Jitt	93,9	0,23	0,79
CPPS 5prc + PPQ	**72,5**	**0,42**	**0,90**
CPPS 5prc + Shim	92,9	0,24	0,80
CPPS 5prc + vAm	93,7	0,23	0,79

Table 3.18:

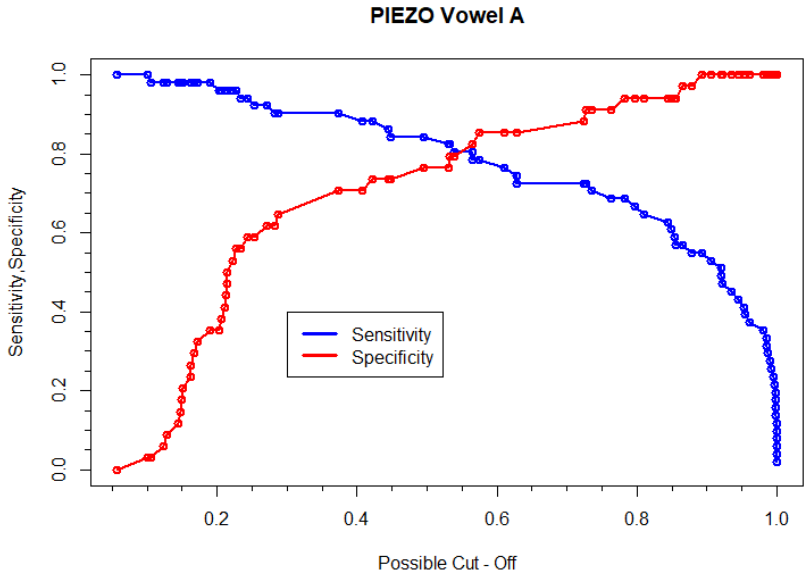


Figure 3.23:

Best model	Int.	Slope 1	Slope 2	Th (P(U))	Sens.	Spec.	Acc.
CPPS 5prc + PPQ	4,2	-0,35	0,12	0,5	0,84	0,77	81%

Table 3.19:

Also for this model, most patients that are under the threshold line (fig. 3.24) have the G of GIRBAS scale equal to 1; three of them have G equal to 2. All the unhealthy subjects with G equal to 3 are above the cut-off. In comparison with the

ECM microphone, the piezoelectric one has a frequency about around 4000-5000 Hz. Consequently, there is the possibility to have more information, because often in pathological voices the disease is visible as a major energy at high frequencies respect to normal voices.

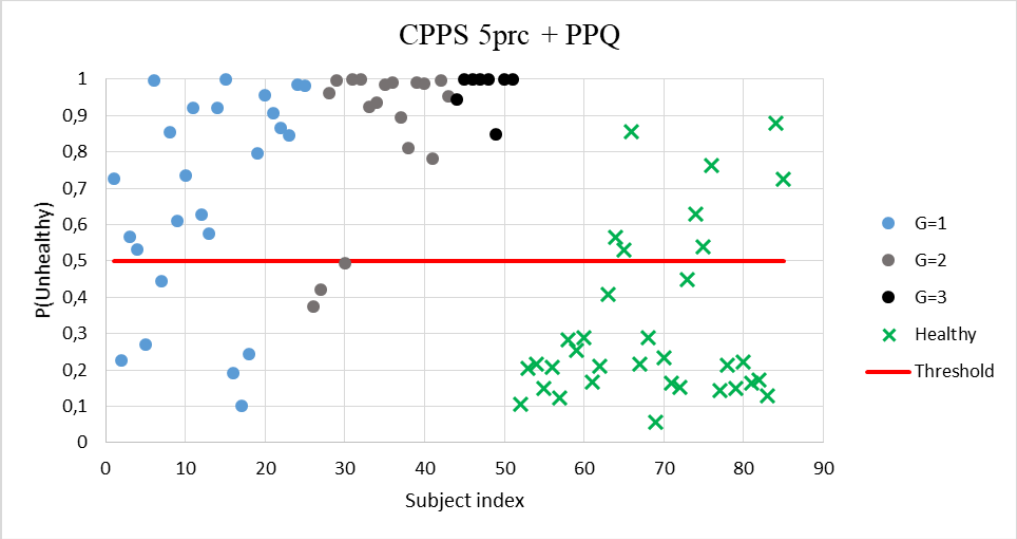


Figure 3.24:

3.4 Continuous speech analysis

3.4.1 Microphone in air: validation of existing models

For Reading and Free speech tasks, detected with the microphone in air, the same procedure illustrated for sustained vowel was followed. Therefore, the starting point of the investigation was related to the results obtained from the antecedent work on the training set; they are shown in table 3.20 and 321, for Reading and Free speech respectively. The instruments of the statistical analysis had allowed to identify the 95^{th} percentile as the best indicator of vocal health status for both Reading and Fee speech, with an AUC of 0.86 in both cases, although the overall results of Reading had been better than the free speech ones. Considering the moderate diagnostic precision, the threshold selection had performed for both

speech materials. The outputs of the two models and the performance indicators are summarized in table 3.22 and 3.23.

Reading

CPPS parameter	AIC	Mcfadden R^2	U-test	AUC
CPPS mean	105,4	0,30	0,000	0,84
CPPS median	108,4	0,27	0,000	0,83
CPPS mode	116,6	0,22	0,000	0,79
CPPS std	127,9	0,14	0,000	0,74
CPPS range	108,4	0,27	0,000	0,83
CPPS 5prc	124,0	0,17	0,000	0,78
CPPS 95prc	**98,9**	**0,34**	**0,000**	**0,86**
CPPS skewness	126,8	0,15	0,000	0,72
CPPS kurtosis	147,3	0,01	0,685	0,48

Table 3.20:

Free speech

CPPS parameter	AIC	Mcfadden R^2	U-test	AUC
CPPS mean	112,10	0,24	0,000	0,80
CPPS median	113,30	0,23	0,000	0,79
CPPS mode	123,10	0,16	0,000	0,79
CPPS std	126,10	0,14	0,000	0,76
CPPS range	118,90	0,19	0,000	0,80
CPPS 5prc	126,60	0,14	0,000	0,74
CPPS 95prc	**99,90**	**0,33**	**0,000**	**0,86**
CPPS skewness	127,50	0,13	0,001	0,70
CPPS kurtosis	145,50	0,01	0,369	0,55

Table 3.21:

CPPS parameter	Int.	Slope	Th (P(U))	Th(dB)	CI (dB)	Sens.	Spec.	Acc.
CPPS 95prc	27,5	-1,52	0,58	18,1	0,6	0,82	0,77	80%

Table 3.22:

CPPS parameter	Int.	Slope	Th (P(U))	Th(dB)	CI (dB)	Sens.	Spec.	Acc.
CPPS 95prc	24,3	-1,4	0,57	17,9	0,6	0,78	0,74	77%

Table 3.23:

In order to perform the validation linked to the generalization ability of the classifier, in this study, the fitted values of the subjects from the test set were calculated for both models, through the respective intercept and slope. Then, according to the threshold value, the same subjects were classified and the Confusion matrix was extracted. Fig. 3.25 and 3.26 report the values of sensitivity, sensibility and accuracy for Reading and Free speech, respectively. Comparing the latter with the values obtained classifying the training set subjects, it is possible to observe that for sensitivity the values are similar, while as regards specificity, the values are lower, in both reading and free speech. This meanings that the two found classifier were not so able to classify correctly new healthy subjects.

CM	Unealthy (22)	Healthy (8)
Positive	19	3
Negative	3	5

$$Accuracy = \frac{19+5}{22+8} * 100 = 77\%$$

$$Sensitivity = \frac{19}{22} * 100 = 81{,}8\%$$

$$Specificity = \frac{5}{8} * 100 = 62{,}5\%$$

Figure 3.20:

CM	Unealthy (24)	Healthy (9)
Positive	19	3
Negative	5	6

$$Accuracy = \frac{19+6}{22+8} * 100 = 75{,}8\%$$

$$Sensitivity = \frac{19}{24} * 100 = 79{,}2\%$$

$$Specificity = \frac{6}{8} * 100 = 66{,}7\%$$

Figure 3.21:

3.4.1 Microphone in air: a new model

The analysis carried on with testing two-variables models; in tab. 3.24 and 3.25 there are the values of Pearson correlation coefficients considering all the possible couples between CPPS parameters, for reading and free speech, respectively. Only the couples that show poor correlation (in yellow) have been used in the statistical analysis.

R	mean	median	mode	std	range	5prctile	95prctile	skew	kurt
mean		0,99	0,89	0,72	0,69	0,85	0,93	-0,95	-0,47
median			0,91	0,69	0,66	0,84	0,90	-0,96	-0,45
mode				0,59	0,56	0,74	0,79	-0,86	-0,37
std					0,79	0,30	0,90	-0,58	-0,76
range						0,44	0,82	-0,56	-0,55
5prctile							0,66	-0,85	-0,16
95prctile								-0,82	-0,65
skewness									0,49
kurtosis									

Table 3.24:

FS	mean	median	mode	std	range	5prctile	95prctile	skew	kurt
mean		0,99	0,89	0,65	0,79	0,87	0,92	-0,95	-0,36
median			0,91	0,63	0,75	0,86	0,89	-0,96	-0,33
mode				0,54	0,65	0,74	0,76	-0,86	-0,29
std					0,71	0,27	0,84	-0,52	-0,79
range						0,60	0,88	-0,66	-0,45
5prctile							0,68	-0,86	0,00
95prctile								-0,80	-0,61
skewness									0,30
kurtosis									

Table 3.25:

The following table (table 3.26 and 3.27) show the results in terms of statistical indicators that express the discrimination power of the tested two-variable models. Only for reading the best model as predictor of voice health status was identified,

and it was the one composed by $CPPS_{range}$ and $CPPS_{mean}$, showing an AUC equal to 0.88, greater than the single-variable model found in the previous study (see table 3.20). By contrast, the analysis did not continue for free speech, since the model with better performance showed results close to $CPPS_{95prc}$ model (equal AUC: 0.88).

Reading			
Two parameters	**AIC**	**Mcfadden R^2**	**AUC**
5prc + std	97,3	0,37	0,87
5prc + 95prc	99,0	0,35	0,87
5prc + range	99,9	0,35	0,87
range + mean	**96,6**	**0,37**	**0,88**
range + median	98,0	0,36	0,88
range + mode	102,1	0,33	0,86
std + median	105,8	0,31	0,85
std + mode	115,4	0,24	0,79

Table 3.26:

Free Speech			
Two parameters	**AIC**	**Mcfadden R^2**	**AUC**
5prc + std	103,9	0,31	0,85
5prc + 95prc	101,8	0,33	0,86
5prc + range	117,0	0,22	0,81
range + mode	116,7	0,22	0,81
std + mean	107,2	0,29	0,84
std + median	108,8	0,28	0,84
std + mode	119,3	0,21	0,79

Table 3.27:

Through the graph depicted in figure 3.22, the cut-off was selected, equal to 0.48 in terms of P(Unhealthy), with a sensitivity of 0.85 and specificity of 0.70. Table 3.28 reports the characteristics and performance of this new classifier and figure 3.23 shows the fitted values for the subject used during the training phase. The threshold divides correctly pathological voices with G equal to 2 and 3 and the

majority with G equal to 1. On the contrary, a relevant number of healthy subjects are under the threshold, but with values near the latter

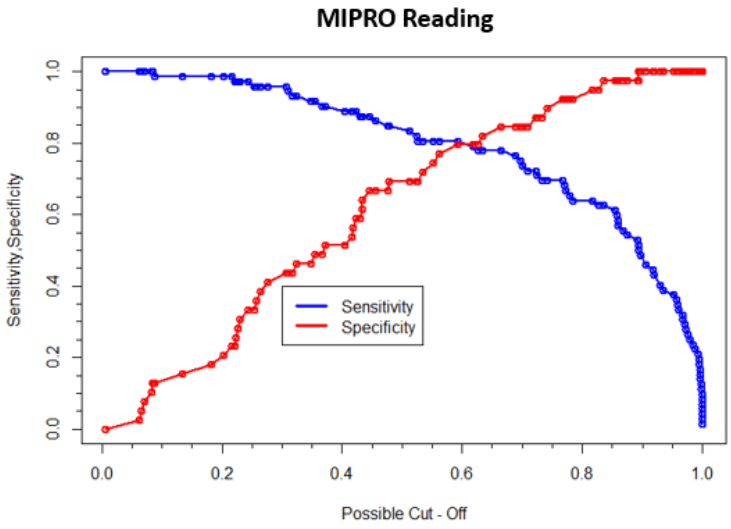


Figure 3.22:

Best model	Int.	Slope 1	Slope 2	Th (P(U))	Sens.	Spec.	Acc.
CPPSrange + CPPSmean	25,5	-0,66	-1,02	0,48	0,85	0,70	80%

Table 3.28:

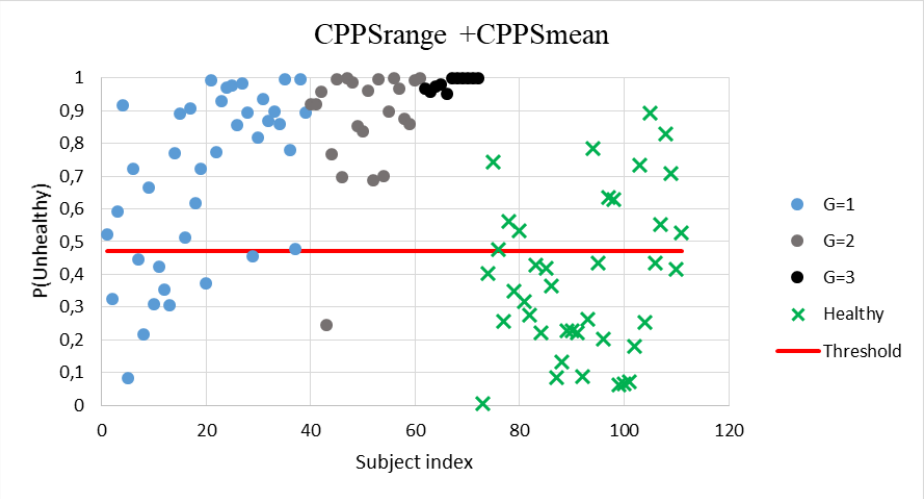


Figure 3.23:

The empirical fitted model for Reading task had the following expression:

$$P(Unhealthy = 1) = \frac{e^{(25{,}5 - 0{,}66 * CPPS_{range} - 1{,}02 * CPPSmean)}}{1 + e^{(25{,}5 - 0{,}66 * CPPS_{range} - 1{,}02 * CPPSmean)}}$$

And it was used for classifying the test set subjects. Confusion Matrix in fig. 3.24 was obtained: sensitivity of 91%, against 85 % of the training set classification, reveals that the model is able, with high precision, to classify new unhealthy subjects; by contrast, specificity, equal to 50 %, has decreased. However the overall accuracy is the same. Such results are visible also in fig. 3.25: also in this case, only patients with G equal to 1 have been wrongly classified or are near the threshold value.

CM	Unhealthy (22)	Healthy (8)
Positive	20	4
Negative	2	4

$$Accuracy = \frac{20 + 2}{22 + 8} * 100 = 80\%$$

$$Sensitivity = \frac{20}{22} * 100 = 90{,}9\%$$

$$Specificity = \frac{4}{8} * 100 = 50\%$$

Figure 3.24:

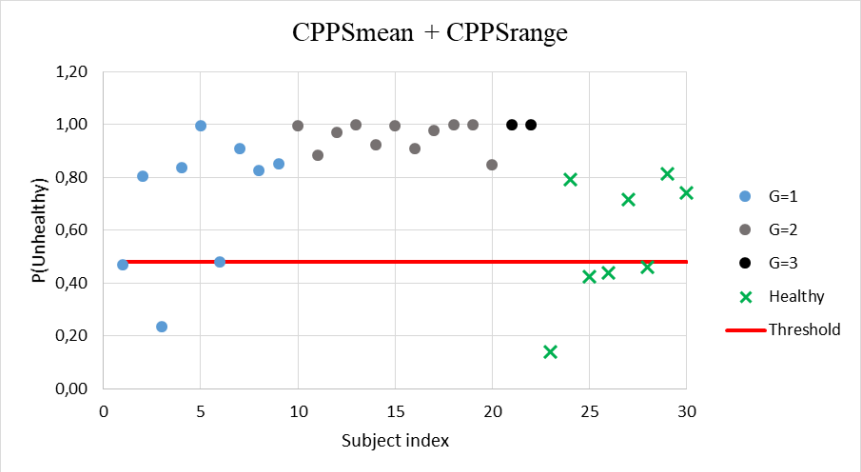


Figure 3.25:

3.4.3 Electret Condenser Microphone: results

In the previous work, unlike the analysis of sustained vowel, the one of continuous speech for ECM microphone had not produced positive results. The logistic regression had not allowed to identify clearly a parameter able to distinguish healthy from pathological voices. For this reason, the subdivision of the data set in training and test set, was not applied in this case, and the whole data set was used in the search for the best model. Table 3.29

4. Voice monitoring of teachers

This chapter wants to investigate the changes of vocal load parameters during working hours in teachers, who represent the main occupational voice users where an excessive and inappropriate use of vocal folds could easily lead to voice diseases. In particular, this study consists in an experimental campaign that is located within a greater system aiming to reduce background noise level in workplaces, in order to ensure comfortable conditions from the vocal fatigue point of view.

In-field experiments have been performed in classrooms during teachers' lessons: on the one hand, long-term monitorings of teachers' voice use have been carried out; on the other hand, a particular detector of background noise, related to external noise and the one produced by students, has been used. Originally, this device was designed not only to measure noise levels, but also to play a semaphore function (SEM) by providing a visual feedback: a green light indicates low noise level, a yellow one middle noise level, and a red one high noise level. According to the light colour emitted, during lesson, students would have to modulate their behaviour so as to reduce the overall background noise. For this purpose, several voice and noise measurements have been made in both conditions SEM-off and SEM-on, at a distance of one month from each other. During data processing phase, several acoustic parameters, as indicators of teacher's vocal load, have been estimated from the SEM-off recordings and the SEM-on one. The aim was to observe an eventual decrease in teachers' vocal load correlated to a decrease in background noise level, confirming all the studies conducted about the Lombard effect, cited in paragraph 2.2. However, this research showed a problem: the vocal load parameters did not seem to improve toward a mild vocal effort and the cause was found in the not reduce background noise during SEM-on monitorings. It is likely that other factors were involved in

noise production, maybe attributable to external noise. For this reason, the data obtained from the signal processing has been used to follow another type of study, namely the analysis of changes in voice production induced by noise level in teachers over a two-month working period and during a working day. In fact, SEM-on recordings were collected after about a month and a half compared to SEM-off ones, and some monitoring related to the same teachers and the same working day have been divided in before and after recreation time. The aim was to observe how teachers' vocal effort changes in the two types of period above mentioned, and thus identify their risk of vocal dysfunctions, evaluating the relationship between vocal load parameters and background noise level. For this purpose, the semaphore device was used as a simple noise level detector, from which L_{A90} distributions (see paragraph 1.4) have been extracted.

4.1 Data collection

4.1.2 Subjects and procedure

Seven voluntary teachers (6 female and 1 male, table 4.1) were involved in the present study, who work in the primary school Roberto D'Azeglio, in the province of Turin (Italy). Their age ranged between 38 and 55 years, with a mean age of 48 years. Physical education teachers were excluded from the study, since they are subjected to a higher vocal effort than science and humanity teachers.

Name	Surname	Gender	Age
Rosalba	Corsentino	F	55
Lidia	Polimeni	F	38
Silvia	Ambiveri	F	47
Carla	Beltramea	F	60
Monica	Mazzei	F	41
Claudio	Calliero	M	47
Rosalinda	Avenia	F	50

Table 4.1: Teachers involved in the experiments.

The vocal activity was monitored during two different time periods, at a distance of about one month and half from each other:

- Stage 1: from 29 January to 2 February 2018
- Stage 2: from 13 March to 22 March 2018

For each stage, two voice monitorings per teacher were performed, but because of several reasons, such as teacher unavailability, problems related to the recording equipment, accidental loss of data, etc. four different recordings were not collected for all teachers. In detail, 12 voice acquisitions were available for stage 1, and 13 ones for stage 2.
Each acquisition included the following four steps, in chronological order:

1. Vowel /a/ scale: from 3 to 5 short vowels /a/ with increasing intensity.
2. Sustained vowel /a/: it is maintained from 3 to 10 seconds.
3. Comfortable speech: teacher speaks freely for about one minute by a comfortable pitch and loudness.
4. Long-term monitoring: it is the vocal recording during working activity, namely during a 3-4 hour lesson.

The first three tasks occurred in a school room as soundproofed as possible, while the long-term monitoring within a classroom.

Before the teacher went to classroom, he had to fill in the PAPV questionnaire (paragraph 3.1.2, fig. 3.2) in order to provide a voice self-evaluation. During working hours, the teacher had to fill in a diary, in which to note the different activities carried out and the relative time bands; while, at the end of the lesson he had to fill in a questionnaire about a self-assessment of the vocal load related to the lesson just occurred. For each voice monitoring, several information were reported, such as the teacher, stage 1 or 2, monitoring number (1or 2), date, morning or afternoon, classroom and the one related to the teacher such as age, gender and experience (years).

4.1.2 Recording equipment and data pre-processing

The recording equipment was made up of:

- A contact Piezoelectric Contact Microphone (PIEZO). It is the same used in the first part of this thesis, related to vocal health, and described in paragraph 3.1.3. It was employed for the recording of all the speech materials, so it is the microphone with which teachers' vocal activity was monitored. Such a contact microphone is suitable for long-term voice monitoring during work activities, since the acquired signal is negligibly affected by background noise, then it is more comfortable than for example an ECM microphone. The device provides a signal expressed in Volts, but one of the vocal load measures is the Sound Pressure Level, in dB. In order to estimate the speech SPL of the signal at a fixed distance d in front of the mouth, a preliminary calibration needs. The latter consists in repeating the vowel /a/ at increasing levels in front of a microphone in air, used as a reference. This is the reason why the first part of the teacher's acquisition includes some vowel /a/ scales: a calibration procedure was needed before starting each monitoring.
- A calibrated Sound Level Meter (SLM, XL2, NTi Audio, Schaan, Liechtenstein) (fig.). It is an omnidirectional and class 1 microphone in air. It was used only in the soundproofed room for vowel scale and sustained vowel. Each subject had to stand in front of the device, on axis. A thin spacer was fixed at a distance of 17cm, between mouth and the sensor. The device samples at 48 kHz and has a resolution of 32 bit. A micro SD card is used to store the acquired data. The signals recorded with this device were used only for the calibration procedure (vowel /a/ scales).
- A device able to measure the background noise activity levels, positioned close to the teacher's desk, at least 1 m away from any reflecting surface and at 1.2 m from the ground, according to the ISO 1996 recommendations []. During the monitoring periods, the classrooms were occupied by an averge numberof 23 students. The background noise level

was evaluated as the A-weighted level exceeded for 90% of the considered time (L_{A90}in dB). All the measurements were performed for a time interval of 5 seconds.

Once the teacher's voice samples were recorded, data was downloaded from the two microphones and saved in a Personal Computer through SD card, as three audio ".wav" files: the first contained the vowel /a/ scales and the sustained vowel, the second the comfortable free speech, and the third the real teacher's voice monitoring during working hours. In this study only the first portion of the first file and the third file were used; the vowel /a/ scale for the calibration procedure, while the voice samples during the lesson time for the tracking of the teacher's vocal behaviour. Before the calibration phase, the vowel scale signals acquired with SLM were resampled at 22050 HZ. Since a typical lesson period consists of various activities with subsequent changes in the voice use and in the noise conditions, a specific activity, i.e. the *plenary lesson*, has been selected to evaluate the occupational voice parameters (OVPs). During this type of lesson, the teacher generally speaks in front of the class with students listening, and only one person speaks at a time.

Therefore, as far as the pre-processing of teaching activity data are concerned, in each monitoring only the time segments containing plenary lesson have been selected for the next analysis, by means of the Audacity 2.2.2 software. This was possible thanks to the teachers' diaries filled in during lesson and listening to each acquisition one at time. Many files were obtained ,with a duration that ranged between 30 min and 80 min, excluding recreation time. Moreover, for all the voice samples, several information were collected: name and surname of the teacher, stage 1 or 2, monitoring number (1 or 2), date, time band, recording duration, in order to simplify the collection of the corresponding noise level values.

4.2 Methods and data processing

4.2.1 Calibration procedure

A script Matlab was implemented on PIEZO and SLM vowel scale signals in order to obtain for each voice monitoring, belonging to a certain teacher and performed in a specific day, the calibration curve. In the algorithm, the samples of the PIEZO signal and the SLM one (resampled) are loaded, grouped into frames of 1024 samples and then processed in order to estimate the root mean square (RMS) values Vpiezo and Vslm for each frame, respectively. After that, in order to select the same time segments containing one or more vowel /a/ scales, a temporal alignment between the two signals is executes. Then, for both PIEZO and SLM signals, the thresholds are manually selected, PIEZO threshold and SLM threshold, so that also the vowels /a/ with very low RMS value are detected. The PIEZO threshold is saved for the next processing. All the values above the threshold are maintained, for both signals, and are used to identify the linear function that relates the PIEZO signal in terms of dB (20*log10(Vpiezo/10^{-3})) to the reference SPLs at the fixed distance from the mouth of the subject under monitoring, $SPL_{ref@17cm}$, obtained through the value of a calibration constant (expressed in Volt/Pa). The interpolating line, i.e. the calibration curve, is plotted (fig.4.1) and the values of intercept and slope are saved for the next analysis of teaching activity signals.

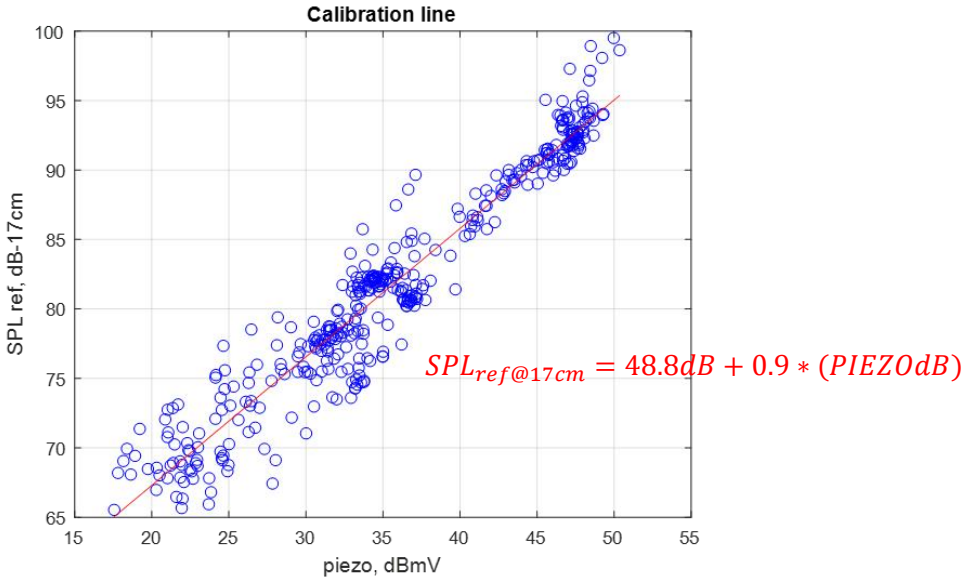


Figure 4.1: Example of calibration curve and respective formula.

4.2.2 Vocal load parameters and background noise level

After the calibration procedure, which is needed before starting each monitoring, the off-line processing allows the following vocal load parameters to be extracted from each plenary lesson signal: the fundamental frequency (F_0 in Hz), the voicing time percentage (Dt% in%) and the sound pressure level at 1m in front of the speaker's mouth (SPL_{1m} in dB). They are estimated through a Matlab script, where the voiced and unvoiced frame detection through a suitable RMS voltage threshold is implemented, considering the signal divided into frames of 1024 samples (about 46ms). The RMS values of each frame is compared to the threshold value (defined during the calibration phase): only for the over-threshold frames, F_0 and SPL_{1m} are calculated; by contrast, zero is assigned to the under-threshold frames. Dt% is obtained by counting how many frame are different from zero in respect to the whole number of frame. Therefore, the Matlab script allows obtaining F_0 and SPL_{1m} occurrence histograms from voiced frames with a bin resolution of 1dB. Mean, median, mode and standard deviation values have been

calculated from such histograms, obtaining $F0_{mean}$, $F0_{median}$, $F0_{mode}$, , $F0_{sd}$, and SPL_{mean}, SPL_{median}, SPL_{mode}, SPL_{sd} respectively. Also the equivalent SPL at 1m from the speaker's mouth $SPL_{eq,1m}$ has been estimated, which express the speaker's vocal effort according to the ANSI S3.5-1997 standard. $SPL_{eq,1m}$ has been calculated as the average of the voiced energy over all the frames, including the unvoiced ones, whose energy is set to zero, according to Svec et al. as follow:

$$SPL_{eq} = 10log\left(\frac{1}{N}\sum_{1=1}^{N} n_i * 10^{\frac{SPL_i}{10}})\right)$$

where N is the total number of frames in the analysed speech and n is equal to 0 for the unvoiced frames and 1 for the voiced frames.

Figure 4.2 and 4.3 illustrate an example of F0 histogram and a SPL one, respectively.

After having calculated all the vocal load parameters and the equivalent SPL as a vocal effort measure for all the plenary lesson recordings, the background noise levels during working hours, at the same time bands, were collected from the noise level detector. Every 5 seconds the L_A90 was estimated as noise level measure, obtaining a distribution, from which mean median and mode were extracted. All the data were organised in Excel putting in evidence, for each acquisition, the stage 1 or 2, before or after the recreation time, Dt%, SPL_{eq}, and the descriptive statistics of L_A90, F0 and SPL distributions. After that, the relationship between the vocal load measures and the background noise was investigated.

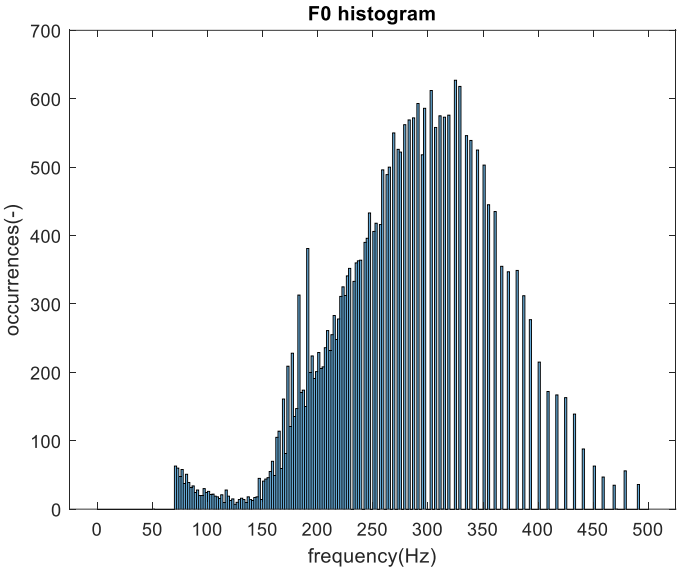


Figure 4.2: Example of F0 histogram.

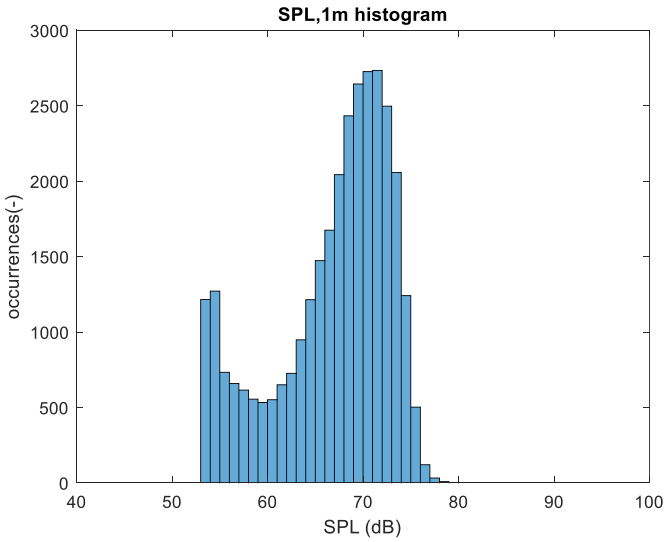


Figure 4.3: Example of SPL histogram.

4.3 Results and discussion

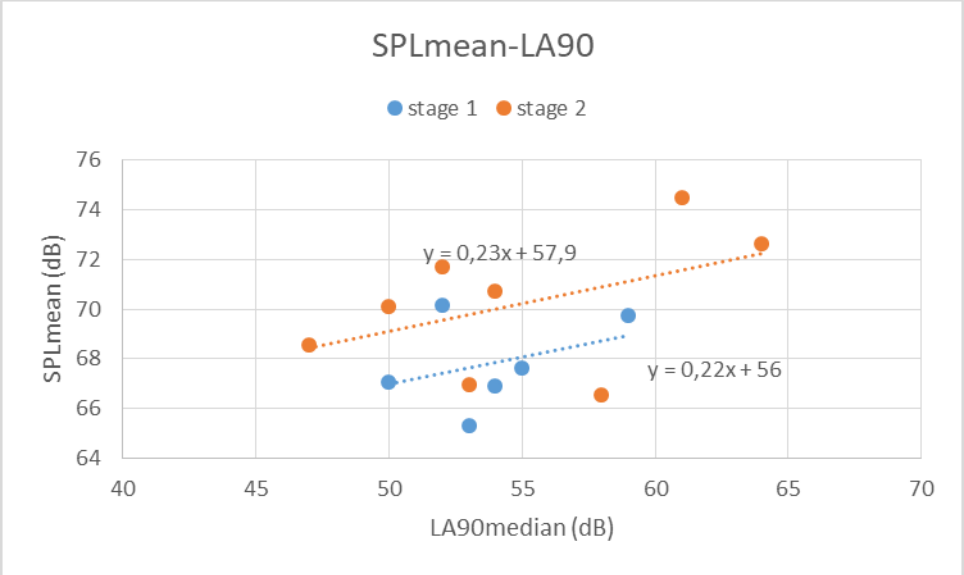


Figure 4.4: Relationship between the mean value of SPL distribution and The median value of LA90 distribution

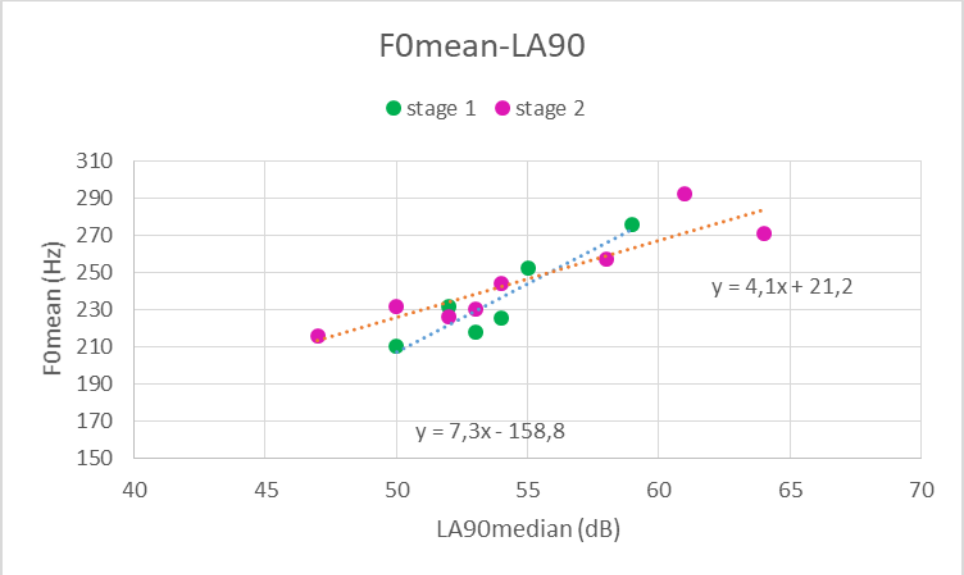


Figure 4.5: Relationship between the mean value of F0 distribution and The median value of LA90 distribution

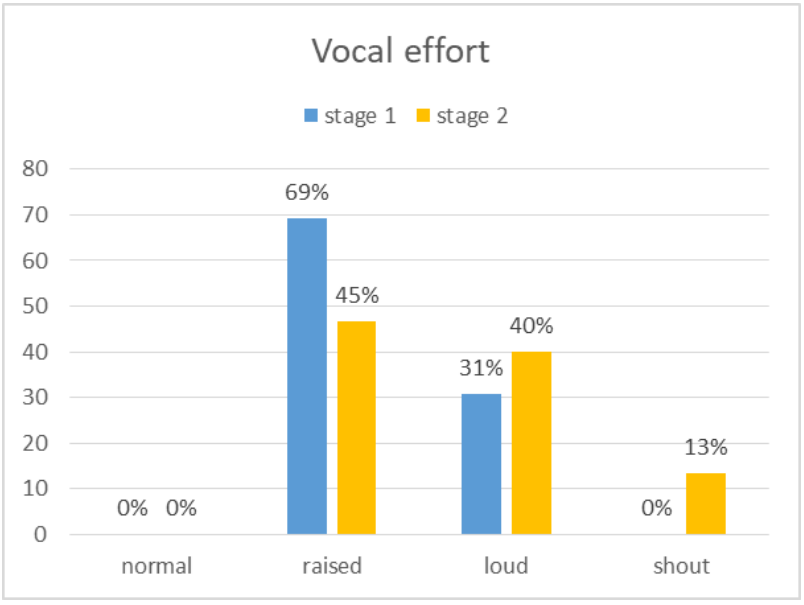


Figure 4.6: Vocal effort evaluation in the two different time periods

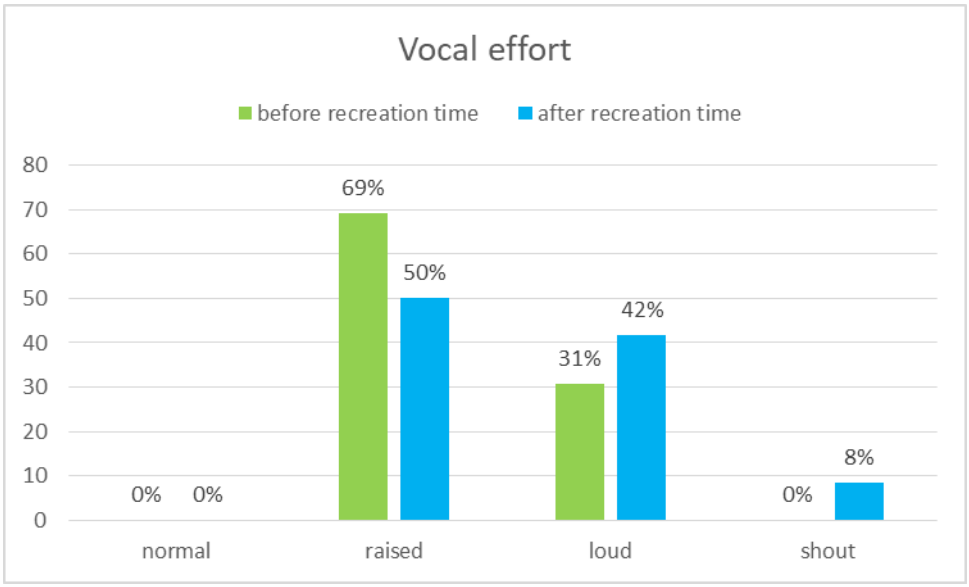


Figure 4.7: Vocal effort evaluation before and after the recreation time.

Bibliografy

[1] W. De Colle, Voce & computer, Analisi acustica digitale del Segnale Verbale (Il sistema CSL-MDVP), Omega Edizioni, 2001.

[2] «Standard, acoustical terminology,» *American National Standards Institute,* vol. S1.1, 1960.

[3] P. David e R. Robert, The handbook of speech perception, John Wiley & Sons, 2008.

[4] E. Vilkman, «Voice problems at work: achallenge for occupational safety and health arrangement,» *Folia Phoniatrica et logopaedica,* vol. 52, n. 1-3, pp. 120-125, 2000.

[5] M. Behlau, F. Zambon, A. C. Guerrieri e R. Nelson, Epidemiology of voice disorders in teachers and nonteachers in brazil: prevalence and adverse effects, Journal of voice, 26(5):665-e9, 2012.

[6] P. Godall, C. Gassull, A. Godoy e M. Amador, «Epidemiological voice health map ofthe teaching population of granollers (barcelona) developed from the eves questionnaire and the vhi.,» *Logopedics Phoniatrics Vocology,* vol. 40, n. 4, pp. 171-178, 2015.

[7] E. Vilkman, «Occupational safety and health aspects of voice and speech professions.,» *Folia Phoniatrica et Logopaedica,* vol. 56, n. 4, pp. 220-253, 2004.

[8] R. Buekers, E. Bierens, H. Kingma e E. Marres, «Vocal load as measured by the voice accumulator,» *Folia Phoniatrica et logopaedica,* vol. 5, n. 47, pp. 252-261, 1995.

[9] S. Whitling, R. Rydell e V. L. Ahlander, «Design of a clinical vocal loading test with long-time measurement of voice,» *Journal of voice,* vol. 2, n. 29, pp. 261-e13, 2015.

[10] I. R. Titze, J. G. Svec e P. P. S, «Vocal dose measures quantifying accumulated vibration exposure in vocal fold tissues,» *Journal of Speech, Language, and Hearing Research,* vol. 4, n. 46, pp. 919-932, 2003.

[11] A. Nacci, B. Fattori, V. Mancini, E. Panicucci, F. Ursino, F. Cartaino e S. Berrettini, «The use and role of the ambulatory phonation monitoe (apm) in voice assessment.,» *ACTA otorhinolaryngologica italica,* vol. 1, n. 33, p. 49, 2013.

[12] International Organization for Standardization, *Ergonomics-Assessment of Speech Communication,* Geneva, Switzerland: ISO 9921, 2003.

[13] C. Manfredi e P. H. Dejonckere, «Voice dosimetry and monitoring, with emphasis on professional voice diseases: Critical review and framework for future research,» *Logopedics Phoniatrics Vocology,* vol. 2, n. 41, pp. 49-65, 2016.

[14] Y. Maryn, P. Corthlas, P. V. Cauwenberge, N. Roy e D. B. Marc, «Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels,» *Journal of voice,* vol. 5, n. 24, pp. 540-555, 2010.

[15] J. P. Teixeira, C. Oliveira e C. Lopes, «Vocal acoustic analysis-jitter, shimmer and hnr parameters,» *Procedia Technology,* vol. 9, pp. 1112-1122, 2013.

[16] E. Yumoto, W. J. Gould e T. Baer, «Harmonics-to-noise ratio as an index of the degree of hoarseness,» *The journal of the Acoustical Society of America,* vol. 6, n. 71, pp. 1544-1550, 1982.

[17] Y. Qui e R. E. Hillman, «Temporal and spectral estimations od harmonics-to noise ratio in human voice signals,» *The Journal of the Acoustical Society of America,* vol. 1, n. 102, pp. 537-543, 1997.

[18] Y. D. Heman-Ackah, D. D. Michael e G. S. Goding, «The relationship between cepstral peak prominence and selected parameters of dysphonia,» *Journal of voice ,* vol. 1, n. 16, pp. 20-27, 2002.

[19] P. Gomez-Vilda, R. Fernandez-Baillo, A. Nieto, F. Diaz, F. J. Fernandez-Camacho, V. Rodellar, A. Alvarez e R. Martnez, «Evaluation of Voice Pathology Based on the Estimation of Vocal Fold Biomechanical Parameters,» *Journal of voice,* vol. 21, n. 4, p. 450476, 2006.

[20] M. Nicastri, G. Chiarella, L. V. Gallo, M. Catalano e E. Cassandro, Multimensional Voice Program (MDVP) and amplitud variation parameters in euphonic adult subjects. Normative study., Catanzaro, Italy, 2006.

[21] S. Y. Lowell, R. H. Colton, R. T. Kelley e Y. Hahn C, «Spectral- and cepstral-based measures during continuous speech: capacity to distinguish dysphonia and consistency within a speaker,» *Journal of voice,* vol. 25, n. 5, pp. e223-e232, 2011.

[22] J. W. Tukey, B. P. Borget e M. Healy, «The frequency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, time series analysis,» 1963.

[23] J. Hillenbrand, R. A. Cleveland e R. L. Erickson, «Acoustic correlates of breathy vocal quality,» *Journal of Speech, Language, and Hearing Research,* vol. 37, n. 4, pp. 769-

778, 1994.

[24] J. Hillenbrand e R. A. Houde, «Acoustic correlates of breathy vocal qualty: dysphonic voices and continuous speech,» *Journal of Speech, Language, and Hearing Research,* vol. 39, n. 2, pp. 311-321, 1996.

[25] Y. Maryn e D. Weenink, «Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index,» *Journal of Voice,* vol. 29, n. 1, pp. 35-43, 2015.

[26] C. Sauder, B. Michelle e T. Eadie, «Predicting voice disorder status from smoothed measures of cepstral peak prominence using Praat and analysis of dysphonia in speech and voice (adsv),» *Journal of Voice,* 2017.

[27] A. Castellana, A. Carullo, S. Corbellini, A. Astolfi, M. Spadola Bisetti e J. Colombini, «Cepstral peak prominence smoothed distribution as discriminator of vocal health in sustained vowel,» *Instrumentation and Measurement Technology Conference (12MTC), 2017 IEEE International,* pp. 1-6, 2017.

[28] A. Castellana, A. Carullo, S. Corbellini e A. Arianna, «Discriminating pathlogical voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel,» *IEE Transaction on Instrumentation and Measurement,* 2017.

[29] A. Castellana, A. Carullo e A. Astolfi, «Dysphonia Recognition in Reading and Free-Speech through Cepstral Peak Prominence Smoothed Estimated by Microphone in Air,» *IEEE Instrumentation and Measurement Technology Conference,* 2018.

[30] L. C. C. Cutiva e A. Burdorf, «Medical costs and productivity costs related to voice symptoms in colombian teachers,» *Journal of Voice,* vol. 29, n. 6, pp. 776-e15, 2015.

[31] J. G. Svec, I. R. Titze e P. S. Popolo, «Estimation of sound pressure levels of voiced speech from skin vibration of the neck,» *The Journal of the Acoustical Society of America,* vol. 3, n. 117, pp. 1386-1394, 2005.

[32] H. Lane e B. Tranel, «The lombard sign and the role of hearing in speech,» *Journal of Speech, Language, and Hearing Research,* vol. 4, n. 14, pp. 677-709, 1971.

[33] P. Bottalico e A. Astolfi, «Investigations into vocal doses and parameters pertaining to primary school teachers in classrooms,» *th Journal of the Acoustical Society of America,* vol. 4, n. 131, pp. 2817-2827, 2012.

[34] G. E. Puglisi, A. Astolfi, L. C. C. Cutiva e A. Carullo, «Four-day-follow-up study on the voice monitoring of primary school teachers: Relationships with conversational task and classroom acoustics,» *The Journal of the Acoustical Society of America,* vol. 1, n.

141, pp. 441-452, 2017.

[35] K. V. Phadke, A.-M. Laukkanen, I. Ilomaki, E. Kankare, E. Geneid e J. G. Svec, «Cepstral and Perceptual Investigations in Female Teachers With Functionally Healthy Voce,» *Journal of voice,* 2018.

[36] N. Roy, R. M. Merrill, S. Thibeault, R. A. Parsa, s. D. Gray e E. M. Smith, «Prevalence of voice disorders in teachers and the general population,» *Journal of Speech, Language, and Hearing Research,* vol. 2, n. 47, pp. 281-293, 2004.

[37] G. Fava, N. P. Paolillo, G. Oliveira e M. Behlau, «Cross-cultural adaptaton of the italian version of the voice activity participation profile,» *CoDAS,* vol. 26, pp. 252-255, 2014.

[38] F. E. Ferrero, R. Lanni e W. De Colle, Primi risultati di uno studio per la validazione del sistema MDVP come strumento per una caratterizzazione multiparametrica della voce, Acta Phon. Lat., 1995.

[39] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonic to noise ratio of a sempled sound, IFA Proceedings 17, 1993.

[40] K. Hajiian-Tilaki, «Receiver operating characteristic (ROC) curve analysis for medical test evaluation,» *Caspian journal of internal medicine,* vol. 4, n. 2, p. 627, 2013.

[41] G. D'Arrigo, F. Provenzano, C. Torino, C. Zoccali e G. Tripepi, «I test diagnostici e l'analisi della curva ROC,» *G Ital Nefrol,* vol. 28, n. 6, pp. 642-647, 2011.