Assessment of vocal cord nodules: a case study in speech processing by using Hilbert-Huang Transform

Assessment of vocal cord nodules: a case study in speech processing by using Hilbert-Huang Transform

M Civera[1], C M Filosi[2], N M Pugno[3,4,5], M Silvestrini[2], C Surace[1], K Worden[6]

[1] Politecnico di Torino, Department of Structural, Building and Geotechnical engineering, Corso Duca degli Abruzzi, 24, 10129, Turin, Italy

[2]Azienda Provinciale per i Servizi Sanitari - Provincia Autonoma di Trento, 38123, Trento, Italy

[3] Laboratory of Bio-Inspired and Graphene Nanomechanics, Department of Civil, Environmental and Mechanical Engineering, University of Trento, 38123, Trento – Italy;

[4]School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, E1 4NS London - United Kingdom;

[5]Italian Space Agency, Via del Politecnico snc, 00133 Rome, Italy.

[6] Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK.

E-mail: marco.civera@studenti.polito.it

Abstract. Vocal cord nodules represent a pathological condition for which the growth of unnatural masses on vocal folds affects the patients. Among other effects, changes in the vocal cords' overall mass and stiffness alter their vibratory behaviour, thus changing the vocal emission generated by them. This causes dysphonia, i.e. abnormalities in the patients' voice, which can be analysed and inspected via audio signals. However, the evaluation of voice condition through speech processing is not a trivial task, as standard methods based on the Fourier Transform, fail to fit the non-stationary nature of vocal signals. In this study, four audio tracks, provided by a volunteer patient, whose vocal fold nodules have been surgically removed, were analysed using a relatively new technique: the Hilbert-Huang Transform (HHT) via Empirical Mode Decomposition (EMD); specifically, by using the CEEMDAN (Complete Ensemble EMD with Adaptive Noise) algorithm. This method has been applied here to speech signals, which were recorded before removal surgery and during convalescence, to investigate specific trends. Possibilities offered by the HHT are exposed, but also some limitations of decomposing the signals into so-called intrinsic mode functions (IMFs) are highlighted. The results of these preliminary studies are intended to be a basis for the development of new viable alternatives to the softwares currently used for the analysis and evaluation of pathological voice.

1. Introduction.

Healthy condition of vocal cords is of paramount importance, as their mucosal vibration is the source of the human vocal emissions [1]. Thus, studies about the vibratory mechanisms of vocal folds have been conducted for over twenty years [2][3], as well as research about the frequency content of the acoustic output generated by them [4]. Both of these are nontrivial tasks, as human speech represents a complex issue, mainly due to its well-known nonlinearity and non-stationarity [5][6]. Also, the acoustics of human speech has been studied and modelled at least since the 1960s [7] and progress is still in development nowadays, especially for the realisation of naturally-sounding synthesised voices [8].

The Discrete Fourier Transform (DFT), which is traditionally the preferred and most used signal processing tool, is limited by its assumption of stationarity of the given signal; in the case of human voice, this supposition does not hold true, as the mean, variance and other statistical parameters change with time [9]. Algorithms currently used for the extraction of fundamental frequencies in voice recordings generally assume the signal to be generated by a linear source (which is not), and to be locally stationary [10]. These assumptions are too oversimplified in the case of pathological vocal emissions, which, even more than normal speech signals, are characterised by nonlinearities due to incomplete closure of the vocal cords and aperiodic vibrations. As a result, alternative methods should be tested, in order to verify their capacity to properly address the topic.

In this study, four voice recordings, originated by an unique patient before and after surgical remove of vocal cord nodules, have been analysed by applying a technique of signal processing - the Hilbert-Huang Transform (HHT) - to Intrinsic Mode Functions (IMFs), obtained through the Complete Ensemble EMD with Adaptive Noise (CEEMDAN). This algorithm has been proposed by Colominas, Schlotthauer and Torres [11] in its current form and applied to electroglottograms, electrocardiograms and electroencephalograms [12]. However, previous versions of the algorithm have been studied by the same authors since 2011 [13]. The EMD has also been applied to the identification of pathological voices and for the classification of patients with Adductor Spasmodic Dysphonia (AdsD) and Muscular Tension Dysphonia (MTD) [14], as a tool for differential diagnosis. Here, the viability of this relatively novel approach is tested for the analysis of healing processes during convalesce and for the estimation of the immediate effects of surgery removal of nodules. In more detail, the HHT disclosed some well-defined trends during the convalescence time, even if it showed also some limitations, mainly due to its being an empirical decomposition. The investigated method could bring to light new non-invasive tools for healing process monitoring, if the identified trends could be systematically observed in a large sample of patients.

The paper is organised as follows: in Section 2, vocal cords nodules are briefly described. Some hint about epistemology, symptoms, signs and treatment are presented, too, in order to provide a global background of this research. An overview of the techniques currently used for the diagnosis concludes this part. In Section 3, the case report is introduced. In Sections 4, the investigated methods, based on Hilbert-Huang Transform, are employed to disclose trends of the frequency response of the given signals; Hilbert Spectra for Time-Frequency Analysis and Marginal Hilbert Spectra for Energy Content Analysis, as obtained from the input signals, are presented and commented. In Section 5 obtained results are shown. Finally, in Section 6 the paper concludes with some overall discussion.

2. Vocal Cords Nodules.

Vocal cords nodules are localised, benign (i.e., noncancerous) and callous-like masses, present on both folds although not necessary symmetrical, and located within the lamina propria. Their onset, development and – if medically treated – removal and/or regression alter the mechanical and vibratory properties of the vocal folds. In most of the cases, no other functions than phonation are compromised; also, for milder cases, lesions may resolve naturally by reducing voice use.

2.1. Causes, Early Signs and Symptoms

Growth of pathological masses in vocal folds is generally caused by excessive and repeated mechanical stress. In fact, vocal cords are subject to collision forces at each vibratory cycle. Moreover, the air forced through the small gap between the folds during voice modulation causes also drying. Therefore, nodules arise from vocal cords tissue trauma, which in turn is due to chronic vocal overuse or misuse. Over time, these vocal abuses generate firstly soft and swollen spots, which then evolve into nodules and become bigger and stiffer if the incorrect vocal use persists.

Ordinarily, the first symptoms noticed by people affected are difficulties to produce sounds belonging to the upper vocal pitch range [15] [16]. Nevertheless, the definition of “healthy” vocal range is still not well defined, and changes according to the field of interest. Conventionally, the limits of “regular” vocal emissions reach as low as 50 Hz (at least) and up to 20 kHz [17]; differences in vocal fold size, due to gender, genetics, age and other causes, makes the definition of “physiological” vocal range quite vague.

However, since nodules interfere more or less markedly – depending on their size – with the vibrational behaviour of the vocal cords, differences between frequency responses of healthy vs. pathological conditions are known and documented in the literature [18], even if the vibrational behaviour of vocal folds is not at all easy to understand nor to reproduce, mainly due to its known nonlinearity [5] [19]. Generally, nodules cause frequency and intensity ranges to be reduced, but fundamental frequencies and intensity often do not undergo any dramatic change when these masses develop [15] [16]. Other prominent signs of vocal fold nodules are breathiness and hoarseness. The latter is a sign due to aperiodic vibrations of vocal fold, while the former results from the incomplete closure of folds upon phonation [15]. The patient's voice may also be perceived as more harsh and rough than usual.

Another common symptom that may arise is a sensation of pain or soreness in the neck, lateral to the larynx. This generally happens because of the increased effort required to produce the voice [15] [16]. As will be discussed in more detail in Sections 4 and 5, this point represents also a great limitation to the usefulness of audio recordings as they are currently performed for vocal cord nodules estimation. Indeed, the patients, more or less involuntarily, arrange their vocal emission in order to provide the requested tone; speech production is a closed loop process – and so patients will adjust in order to produce the correct sound if possible. This means that some parameters, especially volume (and so, energy content of the signal) become meaningless when compared between different audio tracks, as the input (i.e., air pressure as produced by the lungs) is out of control.

2.2. Diagnostic Methods

In the current state of knowledge, a kaleidoscope of techniques is available for the diagnosis of vocal fold nodules. By simplifying, they can be sorted into two main groups: methods that depend on the direct observation of vocal cords and the so-called “non-invasive” ones, based on vocal emissions analysis. It should be remember that “vocal emissions” include all kinds of vocal output – voiced, unvoiced and plosive sounds. Furthermore, non-invasive methods can be divided between the ones based on acoustic measurements – therefore, objective and quantitative approaches – and the ones that rely on perceptual evaluation (a subjective measure of voice, performed by specialists). These two approaches dominate the current state of clinical evaluation of voice quality, not only regarding nodules, but also for all kinds of voice disorders. Even more, perceptual methods can be further divided into clinician-based (e.g., GRBAS/GIRBAS and CAPE-V) and patient-based ones, such as V-RQOL and IPVI. Attempts have been made in order to link objective and subjective estimations [20], while the effectiveness and reliability of subjective tests has been amply discussed in literature [21][22].

In this case, the pre-operative conditions and the follow-ups over convalescence time were investigated through the application of a computerised tool for acoustic voice analysis, the Multi-Dimensional Voice Program (MDVP™). First introduced in 1993, MDVP™ software has been applied in several contexts [23]; its validity has been analysed and compared to other available software in recent years [24]. At least, more than 33 acoustic parameters are currently inspected in a quantitative way thanks to MDVP™ [25]. This has been made possible by the introduction in last decades of new digital instruments, such as the first digital spectrograph (DSP Sonograph), introduced by Kay Elemetrics ® in the late 70s.

Also perceptual evaluation was performed as a first assessment of patient's condition, according to the GIRBAS scale. This method works simply by rating six parameters of the voice, ranging from 1 (non-pathological) to 5 (strongly affected). In its first definition [26], the scale was known as GRBAS, where the five elements considered were the Grade, Roughness, Breathiness, Aesthenia and Strain of vocal emissions. The "I" parameter stands for the Instability of voice and was added at a later stage by [27], making it in its current definition. The GRBAS/GIRBAS scale is currently accepted as standard by the European Group on the Larynx and by the Japanese Society of Logopedics and Phoniatrics and represents the most commonly used perceptual methodology.

2.3. Treatments

In some cases, surgery is not needed, nor recommended, for the treatment of vocal cord nodules [28] [29]. Non-surgical techniques, such as behavioural voice therapy, ordinarily performed by speech-language pathologists, are generally able to produce a reduction in the size and severity of nodules, even if traumatic injuries are unlikely to heal completely without any aftermath [28] [30]. However, removal surgery may be necessary when behavioural interventions are not effective. Nevertheless, it is not impossible that the vocal range will be permanently altered post surgery [31]. In the particular case reported here, the patient underwent a clinical surgery removal, as voice therapy alone proved to be insufficient.

3. Case Report.

The patient, an Italian male adult (one of the Authors), started to suffer from voice disorders in April 2013, probably due to the overuse of voice. The clinical picture showed an upper phlogosis and was mainly characterised by hoarseness worsening as a result of vocal strain; an endoscopic evaluation indicated a haemorrhagic vocal cord polyp and some signs of vocal stress.

Two months later, on 11th June 2013, the patient underwent the removal of the right vocal cord polyp, under analgosedation with remifentanil, according to the technique of Target Controlled Infusion (TCI) and local anaesthesia. The procedure was conducted in an awake setting and the patient was dismissed on the same day.

The patient received one session of preoperative counselling, targeting vocal hygiene instruction and surgery preparation; he also followed a course of postoperative therapy.

On perceptual evaluation, GIRBAS decreased from 1-1-1-1-2-1 to all zeros 5 months postoperatively. A voice handicap index (VHI) was also administered before and 5 months after the operation to evaluate the patients' perceived satisfaction with his voice and it went from mild deficit (score 12) to normal results (score 1 point).

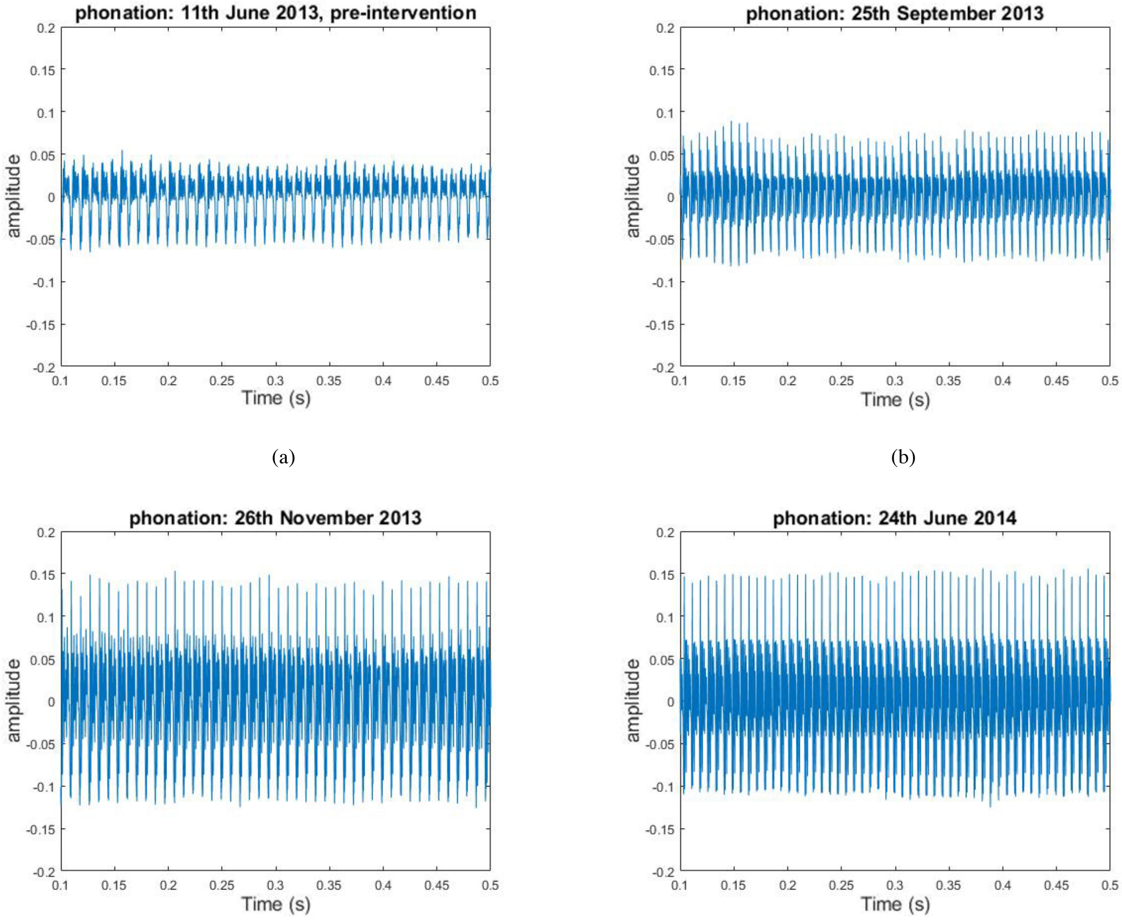
In the acoustic analysis carried out by MDVP™ model, an improvement of the parameters (jitter, shimmer, NHR, high pitch range) were also obtained, as will be explained in more depth later.

Since treatment, the patient has not presented any dysphonia recurrence even if he continues to use the voice intensively, mainly due to his work, which is highly demanding in terms of voice use.

The recordings have been all realised at ENT Unit, Santa Chiara Hospital, Trento, Italy. The microphone and instrumentations are produced by KAYPENTAX, provided by default for MDVP™ analysis.

3.1. Speech Records

Each speech record comes from the same patient, is 3.75 seconds long and is composed by a single emission of a sustained vowel 'a' according to Italian pronunciation (/a/ as defined by the International Phonetic Alphabet). Sampling frequency (i.e., the samples of sounds per second to represent the speech recorded digitally) is 44100 Hz, resulting in 165375 elements inside each digital record. This sampling rate has been set in order to cover the entire 20 – 20000 Hz range of human hearing. The audio tracks have been labelled this way: (I) track #1: 11th June 2013, pre-operative (day of surgery, immediately before operation with pathological voice); (II) track #2: 25th September 2013 (103 days since surgical intervention – i.e. circa three months and a half later); (III) track #3: 26th November 2013 (166 days or circa five months and a half later); (IV) track #4: 24th June 2014 (378 days – i.e. circa one year since surgery; 210 days since previous record). They can be seen in Figure 1. By comparing the first two records, it is possible to have direct insight into the surgical intervention results and the immediate aftermath, while evolution of specific trends between tracks #2, #3 and #4 have also provided knowledge about follow-ups and healing process of voice condition during convalescence.



(a)

(b)

(c)

(d)

Figure 1. Audio tracks of the patient vocal emission (sustained /a/). (a) track #1 (pathological voice). (b) track #2 (circa three months after removal surgery). (c) track #3 (circa five months since operation). (d) track #4 (one year since operation). For all figures, the signal is stopped at 0.5 seconds.

3.2. MDVP Clinical Reports

Four medical reports have been filed from the analysis of all tracks via MDVP™. Among all the parameters considered, average fundamental frequency (*F0*) is the most commonly used for the evaluation of voice disorders; ordinarily, fundamental frequencies will fall between 85 to 180 Hz for male adults, and between 165 to 255 Hz for the same-aged females [32]. *F0* can be automatically tracked by using peak picking strategies, autocorrelation techniques or other equivalent methods [33].

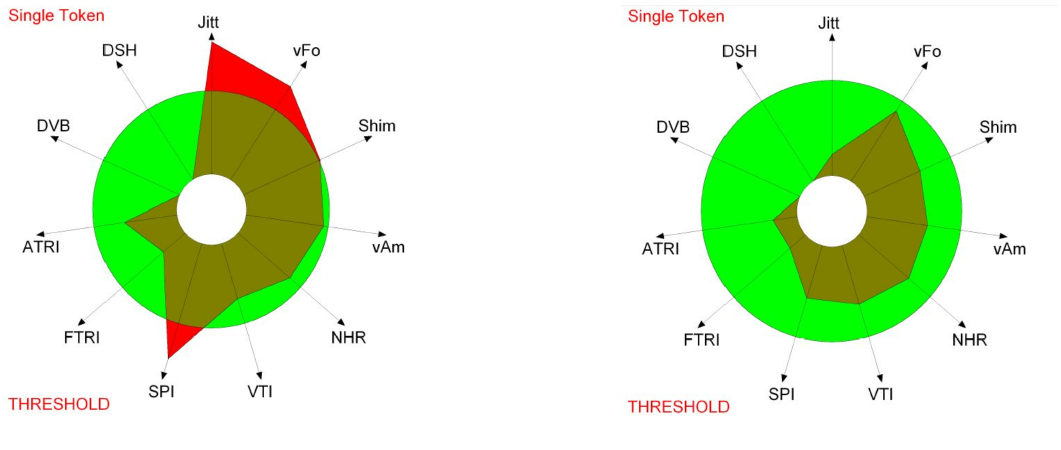
It represents, obviously, the inverse of the fundamental period *T0,* that is to say, the elapsed time between two successive laryngeal pulses [10]. Nevertheless, as stated before, *T0* is defined in a speech signal according to the assumptions of linearity and local stationarity, which can be a problem if used to approximate a pathological voice. As it will be discussed later in this paper, HHT-based frequency related parameters bypass these limitations.

It results from these reports that pathological voice is not subject to any impressive shift in fundamental frequency, as expected ([15] [16]).

In Figure 2, some results from MDVP™ Reports are reproduced. The 11 indicated parameters are (clockwise moving from the top) the Jitter percent (Jitt); the Fundamental Frequency variation (v*F0*); the Shimmer percent (Shim); the Peak-to-Peak Amplitude Variation (vAm); the Noise-to-Harmonic Ratio (NHR); the Voice Perturbation Index (VTI); the Soft Phonation Index (SPI); the F0-Tremor Intensity Index (FTRI); the Amplitude Tremor Intensity Index (ATRI); the Degree of Voice Break (DVB); and the Degree of Sub-Harmonics (DSH). The green circle encloses the threshold of healthy conditions (every parameter has a different scale).

These parameters are not the only ones produced by MDVP™ analysis, but represent those generally most taken into account, as they are considered the most important objective measures for assessment of several voice disorders [23].

On this particular case, the parameters of interest – the ones which exceed the respective thresholds – are the Jitt, the v*F0*, the Shim and vAm. In more detail, v*F0* exceeds its threshold (1.100 %) before removal, but decreases and stabilises since then; Also Jitt falls drastically just after operation (from 1.655 %, with a threshold of 1.040 %, to 0.379 %) and remains inside the limits afterwards. On the other hand, Shim, which is just slightly over the limit before the intervention (3.851 % respect to 3.810 %), increases in the first months of convalescence (to 5.139 %) and then starts to decrease (to 4.003 % and 2.467 %, respectively for track #3 and #4). Also vAm, which was not above the limit before – for pathological conditions –, went beyond the 8.200 % threshold after operation, reaching a maximum of 10.427 % (track #2) and then declining to 6.341 % (track #3) and to 5.255 % (track #4).



(a) (b)

Figure 2. MDVP Clinical Reports. (a) 11th June 2013 (track #1) and (b) 24th June 2014 (track #4).

By considering these data, it is possible to deduce that a transient effect of removal surgery was a temporary increase in variability of the amplitude, which affected both the parameters linked with it, ie. Shimmer and vAm (that is to say, the amplitudes of consecutive periods, divided by the average amplitude, and the Peak-to-Peak Amplitude). Instead, Jitter, which represents the cycle-to-cycle variation of *F0*, and other *F0*-related parameters were positively affected by the surgery and suffered no transient worsening. However, these reports do not provide any information about which frequencies were afflicted the most by the nodules' presence and removal. This information was provided by the EMD-based analysis described here in the next Section.

4. Hilbert-Huang Transform for Speech Processing.

The Hilbert-Huang Transform (HHT) was proposed for the first time in its current form by Huang et al [6]; it represents a suitable option for representing data from nonlinear, non-stationary processes without losing any time-domain information. Essentially, the HHT is made up by two parts: Empirical Mode Decomposition (EMD) and Hilbert Spectral Analysis (HSA). Although the HHT is being used more and more often in the signal-processing context, some background theory will be provided here in order to make this paper a little more self-contained. Much deeper explanations can be found in Huang's original papers [34][6][35], as well as in his recent book [36].

For an arbitrary time series, we can define the Hilbert Transform as

$$HT[g(t)] = \tilde{g}(t) = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{g(\tau)}{(t-\tau)} d\tau \tag{1}$$

This way, the generic time series g(τ) is convoluted with the function 1/πt. This emphasises the local properties of the signal analysed. The tilde ˜ is used here to denote the transformed function, which is still a function of time, as the HT maps function of time or frequency into the same domain, in contrast to the DFT. It should be remarked that the HT, if applied to some dataset, could return physically meaningless results – that is to say, negative frequencies. In order to avoid these problems, two conditions must be applied to the input data: (1) the function must be symmetrical respect to the local zero mean (2) function must have the same number of zero crossing and extrema, or differ at most by one.

Empirical Mode Decomposition provides "modes" that satisfy both these restrictions. These modes – the so-called Intrinsic Mode Functions (IMFs) – can be regarded as the oscillations embedded in the original signal [37]; differently from the harmonics obtained through Fourier Transform, their amplitude and frequency is not constant over time. The process by which they are extracted from the signal can be found in [12]; this step-by-step method is also known as the *sifting process*.

The process is quite straightforward:

(1) for $k = 0$, all the extrema (local maxima and minima) of the analysed data (i.e., of $r_0 = x$) are identified;

(2) local maxima of r_k are connected by a cubic spline line, defining $\mathrm{e_{MAX}}(t)$; likewise, local minima of the same function are linked revolving to the same kind of spline interpolation, thus obtaining $\mathrm{e_{MIN}}(t)$.These two lines form, respectively, the upper and the lower envelops for the given data, which all stand between them;

(3) local mean $\mathrm{m}(t)$ is evaluated as the mean between $\mathrm{e_{MAX}}(t)$ and $\mathrm{e_{MIN}}(t)$;

(4) IMF candidate $\mathrm{d}_{k+1}(t)$ is extracted as $\mathrm{d}_{k+1}(t) = r_k(t) - \mathrm{m}(t)\ \forall t$;

(5) the properties of $\mathrm{d}_{k+1}(t)$ are checked: if $\mathrm{d}_{k+1}(t)$is an IMF, for that instant ' t ' $\mathrm{x}(t)$ is replaced by $\mathrm{d}_{k+1}(t)$, the residue $r_{k+1} = x - \sum_{\mathrm{i}=1}^{k} d_{\mathrm{i}}$ is computed, flag k is increased by one ($k = k + 1$) and r_k is treated as input data for step (2). If $\mathrm{d}_{k+1}(t)$ is not an IMF, $\mathrm{d}_{k+1}(t)$ itself is treated as input data for step (2).

Iteration process ends when the residual r_k satisfies a predefined stopping criterion. In this work, the decomposition has been stopped when, for the *n*-th iteration, the residual $\mathrm{r}_n(t)$ became a monotonic function, from which no more IMFs can be extracted, or when the set number of maximum iterations was reached, whichever came first. Once computed, the IMFs form a complete and nearly orthogonal basis for the original signal; each group is formed by data which have, at any point, zero mean for both the maxima and the minima envelopes [38], as they are – by definition – monocomponent signals.

For this study, the MatLab ® script 'ceemdan.m' has been used. This code has been developed by Marcelo Colominas and was introduced in the current version in [12]. CEEMDAN, or the Complete Ensemble EMD with Adaptive Noise algorithm, is an improvement of the basic EMD algorithm, which is affected by the problem of the so-called "*mode mixing*", the overlap between different modes that have so small differences that they can led to misaddressing of their components. This causes an alias in the frequency-time distribution, leading to a loss of physical meaning of the decomposed data. To solve this issue, Ensemble EMD was first proposed. To keep the discussion brief, the idea is to add white Gaussian noise at each stage of decomposition; then, the generic k-th IMF is computed as the mean over an ensemble of trials ($\overline{IMF_k}$) of the corresponding IMF_k obtained via EMD. Step-by-step, the algorithm can be defined so:
(1) departing from the original signal $x[n]$, $x^i[n]$ is generated, as $x^i[n] = x[n] + w^i[n]$, where $w^i[n]$ are different realisations of white Gaussian noise for $i = (1, \dots, I)$;

(2) each $x^i[n]$ is decomposed by classic EMD, obtaining the modes $IMF_k^i[n]$ for the i-th Gaussian noise and the k-th mode, where $k = (1, \dots, K)$;

(3) k-th IMF is assigned as $\overline{IMF_k}[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_k^i[n]$.

A much more detailed description of the EEMD can be found in [13] and [14]. CEEMDAN, instead, uses each mode for the computation of the next one, sequentially, in a deflationary scheme; basically, in CEEMDAN the several modes are computed as the difference between the current residual and the average of its local means (considering also the noise added to the signal), while in EEMD each $x^i[n]$ is decomposed independently from the other, thus generating I different residuals $r_k^i[n]$ for each mode.

Again, it is possible to describe also the CEEMDAN algorithm in subsequent steps:
(1) I realisation of white Gaussian noise ($w^i[n]$) are used to define $x^i[n] = x[n] + \epsilon_0 w^i[n]$, with ϵ_0 representing the arbitrary value of noise standard deviation for the first step. Then, first modes are computed exactly as for EEMD: $\widetilde{IMF_1}[n] = \overline{IMF_1}[n] = \frac{1}{I}\sum_{i=1}^{I} IMF_1^i[n]$;

(2) for $k = 1$, the first residual is calculated as $r_1[n] = x[n] - \widetilde{IMF_1}[n]$;

(3) departing from the first residual $r_1[n]$, ${r_1}^i[n]$ is generated, as ${r_1}^i[n] = r_1[n] + \epsilon_1 E_1(w^i[n])$, where $w^i[n]$ are different realisations of white Gaussian noise for $i = (1, \dots, I)$ and the operator $E_j(\cdot)$ indicate the whole process that, given a signal, produces the j-th mode by EMD. Then, the second mode is defined as $\widetilde{IMF_2}[n] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_1[n] + \epsilon_1 E_1(w^i[n]))$;

(4) for $k = 2$ onwards, the k-th residuals are computed as $r_k[n] = r_{k-1}[n] - \widetilde{IMF_k}[n]$;

(5) $r_k^i[n] = r_k[n] + \epsilon_k E_k(w^k[n])$ realisations are decomposed. Then, the $(k+1)$-th mode is defined as $\widetilde{IMF}_{k+1}[n] = \frac{1}{I}\sum_{i=1}^{I} E_1(r_k[n] + \epsilon_k E_k(w^i[n]))$;

(6) steps (4) and (5) are reiterated until $k = K$.

The process is relatively time-consuming, as a large number of iterations are generally required, but reduces substantially the risk of mode mixing. In this work, the initial noise standard deviation ϵ_0 has been set to 0.2; the number of realisations (NR) to 15; and the maximum number of sifting iterations allowed to 3000. The code was also required to automatically increase the SNR for every stage.

It is also important to state that all the data from the four audio tracks have been filtered before being decomposed into IMFs. In more detail, a Butterworth low-pass filter of order 10 has been applied in both the forward and reverse directions, to ensure zero-phase distortion. Filtering was needed in order to reduce the effects of background noise from the audio recordings.

From each track, a set of IMFs has been obtained. From these four sets, in order to speed up the comparison process, four subsets have been extracted, considering only IMFs 7 to 11 (Figure 3). The IMFs chosen for the subset have their Mean Frequencies, f_M, close to the estimated fundamental frequencies of each track, as supplied by the MDVP™ clinical reports.

It must be remarked that the four tracks did not provide the same number of IMFs when decomposed, and that this represents a great limitation to the analysed technique, as explained before. In fact, EMD is a purely empirical decomposition – as the name itself states, obviously – and performs blind signal separation. Hence, no physical interpretation can be provided to justify the obtained number of IMFs. In detail, decomposition of track #1 and track #3 produced both 18 Intrinsic Mode Functions, while 19 functions were extracted from track #2 and track #4. In this particular case, the Authors were able to state that IMFs 7 to 11 include, with small differences, the same data for all the tracks; this was only possible by supervising the EMD operations at any iteration, keeping track of the whole process. However, this is a serious limit to any further attempt to generalise the method, as this distinction can be not always enough evident or easy to oversee.

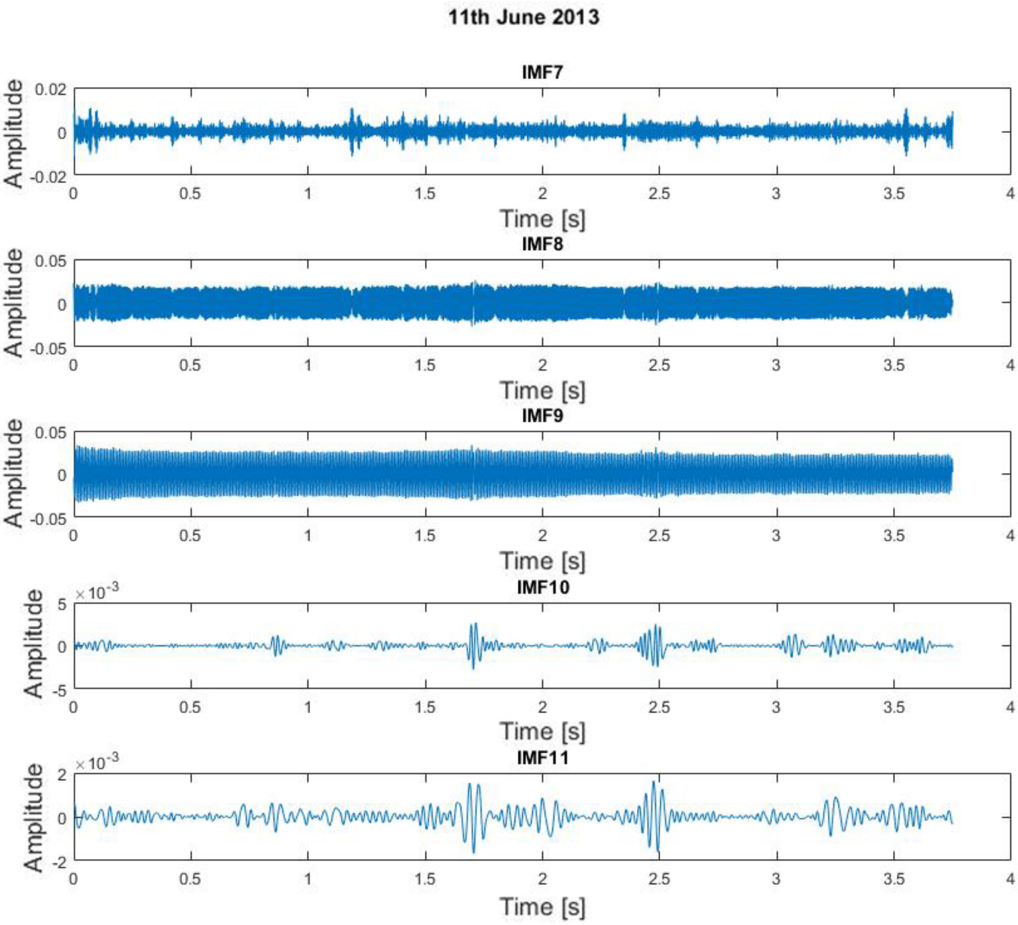


Figure 3. Selected subset (IMFs #7 to #11) for track #1 (11th June 2013).

5. Results.

From each one of the subsets previously described, four parameters have been considered and analysed, in order to test their capability as indices of pathological conditions and/or improvement in healing: the IMF Mean Frequency (f_M), the IMF Standard Deviation (SD), the Total Energy Content (E_{TOT}) and the IMF Content Energy (E_{i}). The first two will be addressed hereinafter as time-frequency parameters, while the latter two as energetic parameters. These features can be defined as follow:

$$f_M[n] = \sum_{\mathrm{i}=1}^{N} f_{\mathrm{i}}\,[n] \tag{3}$$

where N stands for the number of elements inside the investigated signal (here 165375, as the signal is 3.75 seconds long with a sampling rate of 44100 Hz) and $f_{\mathrm{i}}[n]$ are the several instantaneous frequencies of the *n*-th mode.

$$SD[n] = \sqrt{\frac{\sum_{\mathrm{i}=1}^{N}(f_{\mathrm{i}}[n] - f_M[n])^2}{n-1}} \tag{4}$$

where $SD[n]$ simply represents the standard deviation between the several instantaneous frequencies f_{i} and the mean frequency f_M of the *n*-th mode.

$$E_{\mathrm{i}}[n] = \sum_{\mathrm{i}=1}^{N} c^2\,[n] \tag{5}$$

where $c[n]$ is the amplitude, and hence the energy, of the *n*-th mode, computed for any element, as it is not constant in time.

$$E_{\mathrm{TOT}} = \sum_{n=1}^{M} E_{\mathrm{i}}[n] \tag{6}$$

where *M* is the total of the IMFs that compose the original signal (18 for tracks #1 and #3, 19 for tracks #2 and #4).

5.1. The Hilbert Spectrum and Time-Frequency Analysis

The Hilbert Spectrum is the graphical representation of the instantaneous frequency over time, computed separately by applying the Hilbert Transform at each one of the IMFs included into the four subsets defined previously. As one can see in Figure 4, the signal taken before removal surgery shows much more often, high frequency peaks, than in the tracks recorded post-operation.

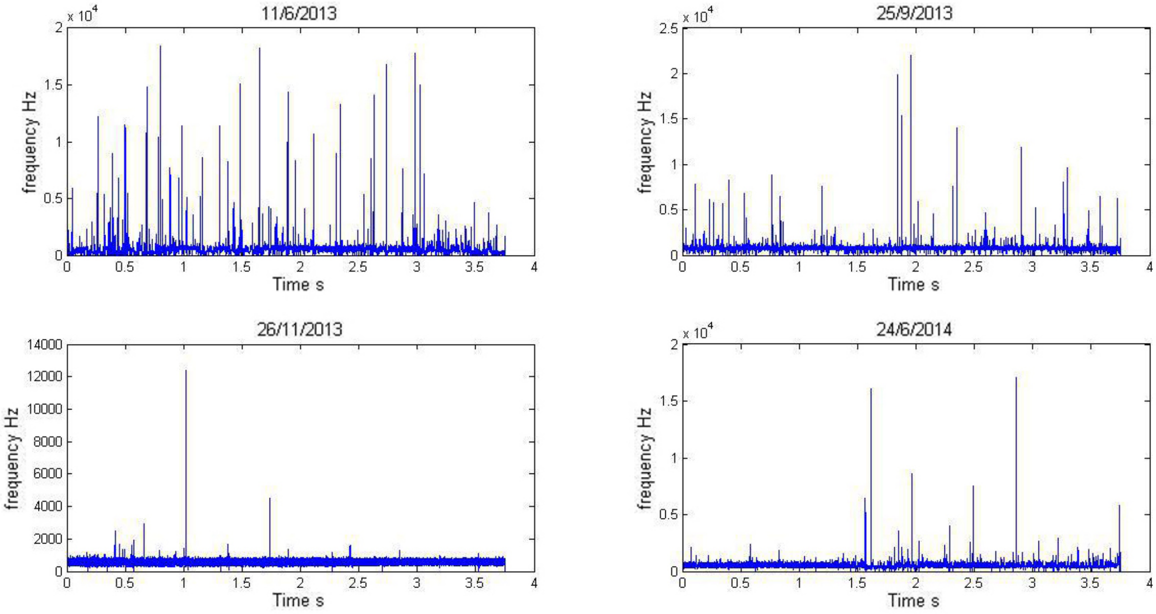
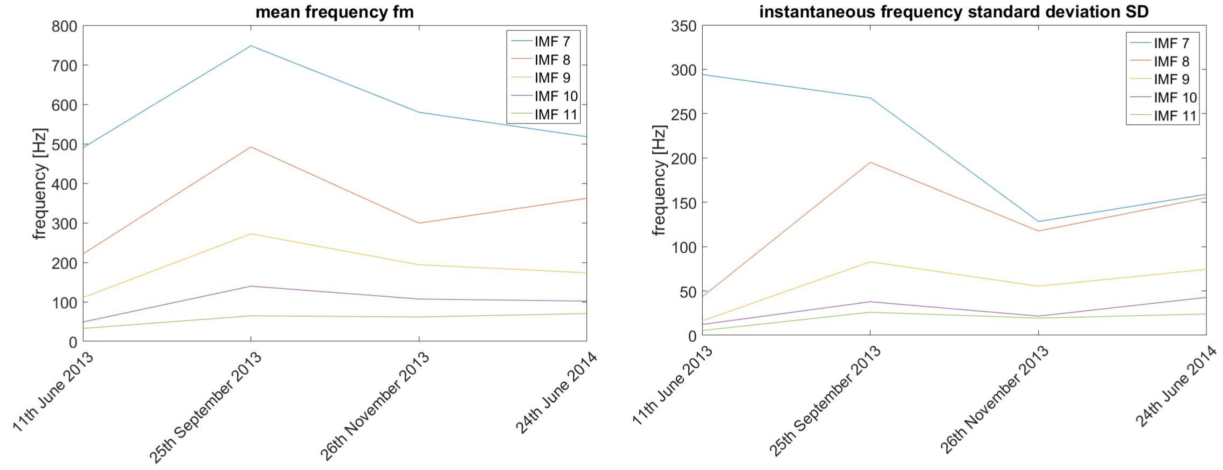


Figure 4. Hilbert Spectra of IMF #7 for all tracks.

By tracking the evolution of both f_M and SD over time, it is possible to recognise some distinct trends. These results, described later, can be seen in the graphs shown in Figure 5. f_M shows a peak corresponding to the post-operative record closest to the surgical intervention (25th September 2013). This reflects the trend of the average fundamental frequency $F0$ between track #1 and track #2, discussed before in Section 3. Nevertheless, $F0$ rebounded slightly after a local minimum occurred corresponding to track #3 and reassessed to values close to the peak for track #4 (mostly the same, circa 134 Hz). Instead, f_M keeps on decreasing for all IMFs, with the sole exception of IMF 8. However, even the 8th mode does not reach again its peak value; overall, all modes seems to tend to stabilising. The exact cause of this behaviour cannot be ascertained with absolute sureness, but could be explained by the fact that surgery is a traumatic event for vocal fold tissues, which change their vibratory characteristics suddenly. A plausible explanation would be scarification and/or swelling. What is certain, is that, since natural frequency is directly proportional to the stiffness and inversely proportional to the mass, any explanation for the immediate post-operative peak will be related to (1) an obvious decrease in mass, due to the nodules' removal; (2) a sudden increase in stiffness, maybe linked to immediate effects of scarification; (3) a combination of these two factors and/or other effects, maybe not related directly to the vocal folds, but to other components of the voice-production mechanisms. Successively, the observed assessment could be most probably due to the healing processes, with scarified tissues reabsorbed over time and a reduction in overall stiffness.

The standard deviation (SD) shows a trend that is strongly related with the one of $F0$. Indeed, since in track #4 the mean frequencies of IMFs 7, 9, 10 and 11 (just to cite the ones reported here) tend to stabilise, while instead $F0$ tends to increase, a higher value of variability is understandable. As mentioned several times earlier, non-pathological voice is supposed to be more coherent and to have a larger range of frequency; both these factors contribute to increase variability of the frequencies contained into the IMFs. Noteworthy is the very high value of SD for IMF 7 in track #1. Most probably, this mode is much more sensitive to the pathological conditions of the pre-operative voice, somehow. A possible explication is that non-homogeneity induced by nodules' presence affects more the modes that are related to higher frequencies, like IMF 7; after surgical intervention, this disturbance is greatly attenuated and this mode starts to behave as its companions do.



(a) (b)

Figure 5. (a) Trends of mean frequencies along time. (b) Trends of instantaneous frequency standard deviations along time.

5.2. Marginal Hilbert Spectra and Energy Content Analysis

Given the Hilbert Spectrum as $H(\omega,t)$ – being, by definition, a portrayal of instantaneous frequency over time – Marginal Hilbert Spectrum can be written as

$$h(\omega) = \int_0^T H(\omega,t) \tag{7}$$

Thus, Marginal Hilbert Spectra (MHS), as shown in Figure 6, permit an immediate visual inspection of the amplitude (i.e., energy) contribution of each frequency.

As can be seen also in Figure 7, each IMF evolves along time according to its own manner, but all of them can be related to the trend of Total Energy E_{TOT}.

The pre-intervention record (track #1) presents a spectrum generally composed of frequencies lower than those preeminent in the post-operative tracks. This evidence agrees with two of the symptoms generally associated with vocal cord nodules, the shrinking of voice range extension and the difficulty to perform the highest frequencies.

It is also possible to notice in Table 1 that no comparison between the IMFs is possible in terms of Energy Content. The values of E_{i} fluctuate from one record to the other in a fashion that gives no noticeable trends. Only two prominent results can be clearly seen here. Firstly, if IMF 11 is excluded, all the other modes seem to reproduce, broadly speaking, the behaviour of the total energetic content, E_{TOT}. As mentioned before, it has been noticed that nodules seem to affect more the IMFs that contain higher frequencies. Thus, IMF 11 should be the least affected of them all; furthermore, its contribution to the E_{TOT} is negligible (always less than 1%), both in general and when compared to the other modes included in the subset. Secondly, it is possible to observe that before surgery, the Energy Content was very different between different modes. These differences results strongly attenuated nine months after the operation; the vertical bands in Figure 10 show this plainly. This can be seen as evidence that, when voice is healed and back to physiological condition, the energy is distributed between modes in a more uniform fashion. However, it has been evaluated that E_{i}, as well as E_{TOT} and any other possible energy-related index, is too affected by variations of the voice volume to produce any reliable result when different tracks are compared; mean frequencies and standard deviations of instantaneous frequencies provided results that look much more reliable.

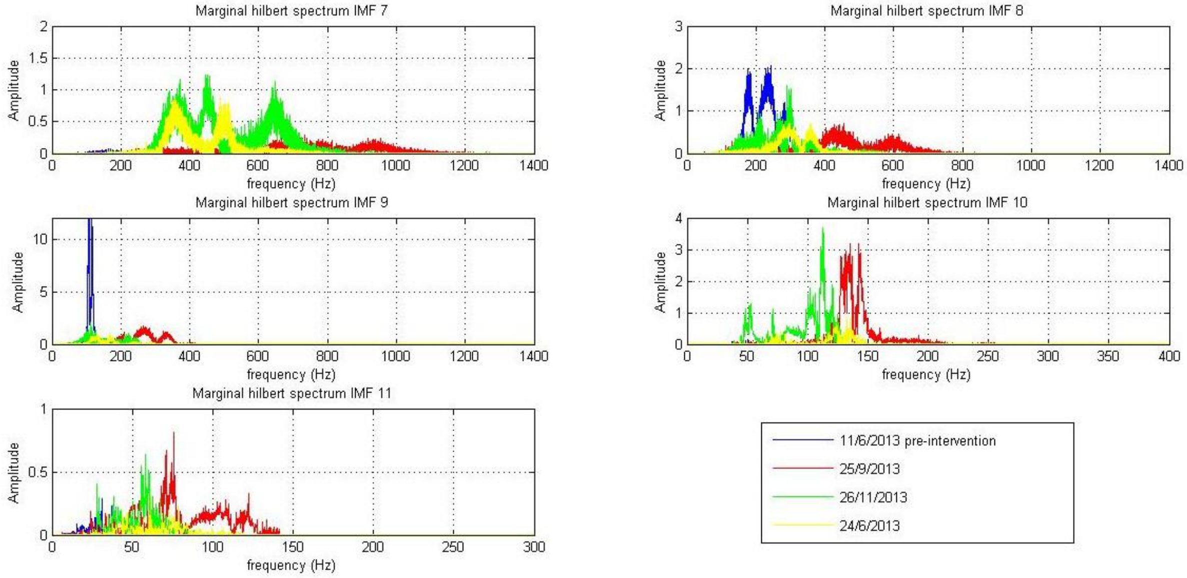


Figure 6. Marginal Hilbert Spectra, IMFs #7 to #11, for all tracks.

Table 1. E_{i} for IMFs 7 to 11, all tracks. E_{TOT} reported for comparison.

	Track #1 (11th June 2013)	Track #2 (25th September 2013)	Track #3 (26th November 2013)	Track #4 (24th June 2014)
IMF 7 [-]	0.970	12.809	54.026	19.361
IMF 8 [-]	25.583	22.236	60.544	7.301
IMF 9 [-]	50.930	32.406	36.188	10.488
IMF 10 [-]	0.053	39.246	25.023	3.133
IMF 11 [-]	0.032	1.22	0.578	0.186
E_{TOT} [-]	88.795	139.972	477.692	312.09

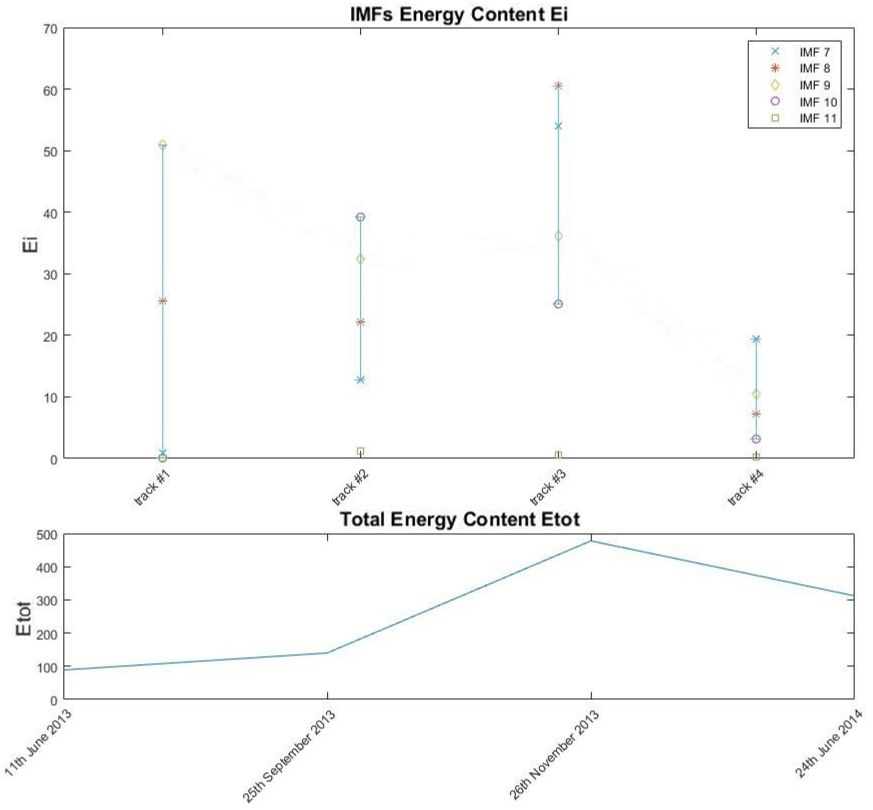


Figure 7. Trends of IMFs Energy Content (Total Energy Content shown for comparison).

6. Conclusions.

The main aim of this research was to investigate the viability of the HHT and CEEMDAN algorithms to track the healing process of vocal cord tissues over convalescence time, as well as to define the immediate aftermath of the removal surgery. A subset has been extracted by the decomposition of the audio tracks into IMFs, specifically IMFs 7 to 11; evolution of these five modes has been studied along the four audio tracks provided. In particular, four parameters have been investigated: the IMF Mean Frequency (f_M), the IMF instantaneous frequency Standard Deviation (SD), the Total Energy Content (E_{TOT}) and the IMFs Content Energy (E_{i}). Some speculations have been drawn from the observed results.

Regarding the trend of f_M over time, all IMFs produced similar result: peak values just after surgery, followed by a stabilisation to values not distant from the pre-operative ones. This proves that vocal cords nodules do not affect substantially the frequency content of voice, as expected, while surgery does, even if for a limited period of time. Higher-numbered modes, which are related to lower frequencies, show to be less affected by the surgery in the short run. Post-operative peaks can be related to a decrease in mass, due to the nodules' removal; to an increase in stiffness, due to tissue scarification; or to a combination of both these effects plus some other mechanisms. The assessment of f_M in the long run, after one year from surgery removal, is almost surely due to an overall reduction of stiffness, which can be related to the healing of scars.

IMFs 8, 9, 10 and 11 showed a higher value of Standard Deviation SD in healed conditions than in pathological ones. This increment is sensibly less marked for higher-numbered modes, as they show again to be less influenced by both nodules' presence and surgery aftermaths. This larger variability of the instantaneous frequencies could be a consequence of the broader range of frequency of the healed voice. IMF 7 counterintuitive behaviour may be a result of the non-homogeneous effects of nodules, as lower-numbered modes proved to be more affected and this "deviation" from the otherwise regular trend is limited only to the pre-operative first record.

It was observed that the range of E_i values among the five IMFs reduced noticeably after convalescence, demonstrating how the energy is more uniformly distributed between modes in healthy vocal conditions. Moreover, the Marginal Hilbert Spectrum of track #1 showed also a composition made mostly by frequencies much lower than those prominent in the following records, thus highlighting the reduction of vocal range to lower-than-usual frequencies that ordinarily occurs with the onset of vocal fold nodules. However, apart these two speculations, energy-related features were considered not to be reliable. Overall amplitude of the four signals resulted to be too heavy influenced by voice volume, a parameter that was not taken in account during the several recording operation. Thus, E_TOT and E_i cannot be completely trustworthy parameters. By comparing SD (which rebounds slightly or stabilises for all the investigated IMFs between track #3 and track #4) with E_TOT and E_i, which instead decrease in all cases, it is also possible to state that removing nodule masses from the vocal cords allows them to vibrate much freely, producing a wider range of frequencies, spending less energy.

To sum up, some points become clear from the investigation performed:

(1) The Hilbert-Huang Transform works, as expected, for the analysis of non-stationary data originated by a nonlinear source, as in the case of human voice; theoretically speaking, the tool is perfectly suited for the non-invasive analysis of vocal fold conditions through speech processing. Different to other current methodologies, the HHT performs time-frequency analysis without requiring any assumption of stationarity of the signal or linearity of the system, thus avoiding the risks inherent in the oversimplifications that other techniques are restrained by.

(2) The EMD-based approach allows analysis of the different frequencies separately, even if the decomposition itself, being empirical, is not free from issues. Since there are no guarantees that the decomposition will produce the same amount of IMFs, if the content of this Mode Functions is not – by chance – clearly similar, as in this particular case, the Hilbert-Huang Transform would be itself viable, but results would be much more difficult, or even impossible, to compare.

(3) Energy Content, both for the overall signal (E_TOT) and for each one IMF inspected (E_i), is too much influenced by the volume of the patient's voice. Some information can nonetheless be extracted from their trends, but it seems that this results must be handled with great care and not uncritically.

(4) IMF mean frequency (f_M) and instantaneous frequency standard deviation (SD) showed clear trends, both during post-operative convalescence time (long run, tracks #2, #3 and #4) and immediately after surgery (short run, tracks #1 and #2). These results are by far the most interesting ones and explain all the possibilities offered by time-frequency analysis for health monitoring the follow-ups after surgical intervention through speech processing.

With all the difficulties and limitations encountered, the authors find that the HHT proved to be feasible, technically speaking, for the proposed aim. f_M and SD can be easily included as additional parameters to MDVP™or to any other software for Acoustic Voice Analysis. Nevertheless, the technical obstacles given by the need to have comparable IMFs between different tracks may be just too extended to make the method advantageous in economical terms. Even more, all the results should be identified in a large, statistically valid population before being accepted definitely. Comparing them with records of patients affected by other voice disorders will also help to discern if the patterns encountered are typical of the pathology studied or detectable for other kinds of disease, too. Even so, the results are encouraging, as trends in time-frequency analysis are evident, and exhort the authors to test other options, especially Wavelets, which could probably overcome the technical problematics that afflict HHT and EMD and provide more crystalline answers.

Acknowledgements

The Authors would like to acknowledge Mr. Jacopo Fiorino, M.Sc. in Biomedical Engineering, for his help in setting the basis of this work. N. M. Pugno is supported by the European Research Council PoC 2015 "Silkene" No. 693670, by the European Commission H2020 under the Graphene Flagship Core 1 No. 696656 (WP14 "Polymer Nanocomposites") and under the Fet Proactive "Neurofibres" No. 732344.

References

[1] Timcke R, von Leden H and Moore P 1958 Laryngeal vibrations: measurements of the glottic wave: part I. The normal vibratory cycle *AMA arch. of otolaryngol.* **68(1)** 1-19

[2] Berke G S and Gerratt B R 1993 Laryngeal biomechanics: an overview of mucosal wave mechanics *J. Voice* **7(2)** 123-128

[3] Titze I R, Jiang J J and Hsiao T Y 1993 Measurement of mucosal wave propagation and vertical phase difference in vocal fold vibration *Ann. Otol. Rhinol. Laryngol.* **102(1)** 58-63

[4] Kasuya H *et al* 1986 Normalized noise energy as an acoustic measure to evaluate pathologic voice *J. Acoust. Soc. Am.* **80(5)** 1329-34

[5] Giovanni A *et al* 1999 Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal folds *J. Voice* **13(4)** 465-476

[6] Huang N E *et al* 1998 The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis *Proc. Math. Phys. Eng. Sci.* **454(1971)** 903-995

[7] Davis S B 1979 Acoustic characteristics of normal and pathological voices *Speech and language: advances in basic research and practice* **1** 271-335

[8] Zen H and Senior A 2014 Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 3844-48

[9] Henríquez P *et al* 2009 Characterization of healthy and pathological voice through measures based on nonlinear dynamics *IEEE transactions on audio, speech, and language processing (TASLP)* **17(6)** 1186-95

[10] Hess W J 2008 Pitch and voicing determination of speech with an extension toward music signals *Springer Handbook of Speech Processing,* ed Springer Berlin Heidelberg pp 181-212

[11] Colominas M A *et al* 2012 Noise-assisted EMD methods in action *Advances in Adaptive Data Analysis* **4(4)** 1250025-1–1250025-11

[12] Colominas M A, Schlotthauer G and Torres M E 2014 Improved complete ensemble EMD: a suitable tool for biomedical signal processing *Biomed. Signal Proces.* **14** 19-29

[13] Torres M E *et al* 2011 A complete ensemble empirical mode decomposition with adaptive noise *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4144-47

[14] Schlotthauer G, Torres M E and Rufiner H L 2010 Pathological voice analysis and classification based on empirical mode decomposition *Development of multimodal interfaces: active listening and synchrony,* ed Springer Berlin Heidelberg pp 364-381

[15] Verdolini K, Rosen C A and Branski R C 2014 Vocal Fold Nodules (Nodes, Singer's Nodes, Screamer's Nodes". *Classification Manual for Voice Disorders-I*, ed Psychology Press pp 37–40

[16] Prater R J and Swift R 1984 *Manual of voice therapy,* ed Little Brown and Company

[17] McKinney J C 1994 *The diagnosis and correction of vocal faults: a manual for teachers of singing and for choir directors (Revised ed.),* ed Genevox Music Group

[18] Deliyski D D 1993 Acoustic model and evaluation of pathological voice production *Eurospeech 1993* **93** 1969-72

[19] Jiang J J, Zhang Y and Stern J 2001 Modeling of chaotic vibrations in symmetric vocal folds *J. Acoust. Soc. Am.* **110(4)** 2120-28

[20] Bhuta T, Patrick L and Garnett J D 2004 Perceptual evaluation of voice quality and its correlation with acoustic measurements *J. Voice* **18(3)** 299-304

[21] Dejonckere P H *et al* 1993 Perceptual evaluation of dysphonia: reliability and relevance *Folia Phoniatrica et Logopaedica* **45(2)** 76-83

[22] Nemr K *et al* 2012 GRBAS and Cape-V scales: high reliability and consensus when applied at different times *J. Voice* **26(6)** 812-e17–812-e22

[23] Christmann M K *et al* 2015 Use of the program MDVP in different contexts: a literature review *Revista CEFAC* **17(4)** 1341-1349

[24] Manfredi C *et al* 2012 Perturbation measurements in highly irregular voice signals: performances/validity of analysis software tools *Biomed. signal proces.* **7(4)** 409-416

[25] Nicastri M *et al* 2004 Multidimensional Voice Program (MDVP) and amplitude variation parameters in euphonic adult subjects. Normative study *Acta Otorhinolaryngol. Ital.* **24(6)** 337-341

[26] Hirano M 1981 Psycho-acoustic evaluation of voice: GRBAS scale. Clinical examination of voice. ed Springer Verlag

[27] Dejonckere P H *et al* 1995 Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements *Revue de laryngologie-otologie-rhinologie* **117(3)** 219-224.

[28] McCrory E 2001 Voice therapy outcomes in vocal fold nodules: a retrospective audit *Int. J. Lang. Comm. Disord.* **36(sup 1)** 19-24.

[29] Pedersen M and McGlashan J 2001 Surgical versus non-surgical interventions for vocal cord nodules *Cochrane Libr.* 1-12

[30] Stemple J C and Hapner E R 2014 *Voice therapy: clinical case studies,* ed Plural Publishing

[31] Benninger M S *et al* 1996 Vocal fold scarring: current concepts and management *Otolaryng. Head Neck* **115(5)** 474-482

[32] Titze I R and Daniel W M 1998 Principles of voice production *J. Acoust. Soc. Am.* **104(3)** 1148-48

[33] Gerhard D 2003 Pitch extraction and fundamental frequency: history and current techniques *Tech. Rep., Department of Computer Science, University of Regina, Regina, Canada* 0-22

[34] Huang N E, Long S R and Shen Z 1996 The mechanism for frequency downshift in nonlinear wave evolution *Adv. Appl. Mech.* **(32)** 59-117C

[35] Huang N E, Shen Z and Long S R 1999 A new view of nonlinear water waves: the Hilbert Spectrum 1. *Annu. Rev. Fluid Mech.* **31(1)** 417-457

[36] Huang N E 2014 Hilbert-Huang transform and its applications, ed World Scientific

[37] Kizhner S *et al* 2004 On the Hilbert-Huang transform data processing system development *2004 IEEE Aerospace Conference Proceedings* **(3)** 1961-79

[38] Fontugne R *et al* 2012 Empirical mode decomposition for intrinsic-relationship extraction in large sensor deployments *Workshop on Internet of Things Applications, IoT-App* **(12)**