Pathological Voice Analysis and Classification Based on Empirical Mode Decomposition

Gastón Schlotthauer[1,4], María E. Torres[1,3,4], and Hugo L. Rufiner[2,3,4]

[1] Lab. de Señales y Dinámicas no Lineales, Fac. de Ingeniería
Universidad Nacional de Entre Ríos, Oro Verde, Entre Ríos, Argentina.
metorres@santafe-conicet.gov.ar

[2] Lab. de Cibernética, Fac. de Ingeniería, UNER, Oro Verde, Entre Ríos, Argentina.

[3] *SINC*(*i*), Fac. de Ing. y Cs. Hs., Univ. Nac. del Litoral, Santa Fe, Argentina.

[4] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina.

Abstract. Empirical mode decomposition (EMD) is an algorithm for signal analysis recently introduced by Huang. It is a completely data-driven non-linear method for the decomposition of a signal into AM - FM components. In this paper two new EMD-based methods for the analysis and classification of pathological voices are presented. They are applied to speech signals corresponding to real and simulated sustained vowels. We first introduce a method that allows the robust extraction of the fundamental frequency of sustained vowels. Its determination is crucial for pathological voice analysis and diagnosis. This new method is based on the ensemble empirical mode decomposition (EEMD) algorithm and its performance is compared with others from the state of the art. As a second EMD-based tool, we explore spectral properties of the intrinsic mode functions and apply them to the classification of normal and pathological sustained vowels. We show that just using a basic pattern classification algorithm, the selected spectral features of only three modes are enough to discriminate between normal and pathological voices.

1 Introduction

Empirical Mode Decomposition (EMD) has been recently proposed by Huang et al. [1] to adaptively decompose nonlinear and non stationary signals in a sum of well-behaved AM-FM components, called Intrinsic Mode Functions (IMFs). This new technique has received the attention of the scientific community, both in applications [2,3] and in its interpretation [4,5]. The method consists in a local and fully data-driven splitting of a signal, in fast and slow oscillations. The advantage of an AM-FM resonance model of speech was previously discussed in [6], where using a Gabor filter bank with six fixed band pass filters, nonlinear features were extracted for instantaneous frequency estimation, phoneme classification, and automatic speech recognition. In this work, we propose two new methods based on EMD (and its variants) with focus in two pathological voice applications: differential diagnosis and fundamental frequency extraction. Preliminary versions of these algorithms were presented in [7,8,9].

The fundamental period T_0 of a voiced speech signal can be defined as the elapsed time between two successive laryngeal pulses, and the fundamental frequency or pitch is $F_0 = 1/T_0$ [10]. Even if F_0 is useful for a wide range of applications, its reliable estimation is still considered one of the most difficult tasks. In speech, F_0 variations contribute to prosody, and in tonal languages, they also help to distinguish segmental categories. Current applications are related with speech and speaker recognition, speech based emotions classifications, voice morphing and the analysis of pathological voices.

In the clinical evaluation of disordered voices, the analysis of F_0 perturbation is a standard procedure in order to assess the severity of pathologies and in monitoring the patient progress [11]. A reliable and accurate estimation of F_0 is essential for this application. Conventional F_0 extraction algorithms assume that speech is produced by a linear system and that speech signals are locally stationary [10]. However, in the case of pathological voices, these assumptions are over-simplifications.

In voice pathology assessment, several parameters extracted from pitch estimation are commonly used. It is therefore important to have a good and reliable F_0 estimation. Unfortunately, no previous method for F_0 extraction operates consistently in the case of pathological voices. This is due to the fact that the vocal folds vibrations of pathological voices present more serious and complex irregularities than the case of normal voices. Some of the difficulties that arise in F_0 estimation, especially when pathological voices are analyzed, include period-doubling and period-halving.

A few EMD based algorithms have been proposed for F_0 extraction [12,13], however they suffer the “mode mixing” problem. Wu and Huang [14] proposed a modification of the EMD algorithm, called Ensemble EMD (EEMD), which largely alleviates this effect, but at the price of a very high computational cost. Taking advantage of its benefits, here we present a new method based on EEMD which is able to extract the F_0 in normal and pathological sustained vowels, improving the behavior of the previous estimators.

In the present paper, we also explore some spectral properties of the IMFs. The comparison of real data IMFs spectra allows us to present preliminary results of an application of this method to the analysis and discrimination between normal and pathological speech signals.

We study a couple of dysphonias with different etiology [15], frequently confused and not easily identified by local clinicians: Adductor Spasmodic Dysphonia (AdSD) and Muscular Tension Dysphonia (MTD).

In recent years, the use of acoustical measures, in combination with pattern recognition techniques, has motivated the appearance of several works concerning the automatic discrimination between pathological and normal voices. In [16], a database with 89 records of the sustained vowel /a/ corresponding to normal and pathological (MTD and AdSD) cases were separated into three classes with a 93.26 % of correct classifications, and into two classes (normal and pathological) reaching a 98.94 %, overcoming the best reported results in the literature. The authors used a pattern recognition scheme with eight acoustical parameters and neural networks. In this paper we show that the spectral properties of the IMFs could be useful to discriminate between normal and pathological voices. These preliminary results suggest that they might provide also clues in order to differentiate between AdSD and MTD.

The paper is organized as follows. In Sec. 2 the data used for the experiments and basic concepts to be used are described. In Sec. 3 the EEMD based F_0 extraction method is presented. In Sect. 4 the pathological voice classification problem is stated and a method based on EMD is described. In Sect. 5 the results for both methods are shown. Finally, in Sec. 6 the discussion and conclusions are presented.

2 Materials and Methods

2.1 Artificial Data

In order to explore the performances of the proposed techniques, experiments were performed with synthetic normal and pathological voices. These signals have been generated using a phonation model that incorporates the perturbations involved in normal voices and in common laryngeal pathologies. This allowed us to maintain controlled experimental conditions, making possible the discussion of the technique and the selection of the appropriate parameters.

The speech signal $y[n]$ was modeled using the classical linear prediction model $y[n] = -\sum_{p=1}^{P} y[n-p]a[p] + x[n]$, where $a[p]$ are the linear predictor coefficients, and $x[n]$ is the input representing the glottal pulses. The input is modeled by a train of pulses, with variable period and amplitude:

$$x[n] = \sum_{k=1}^{K} G[k]\,\delta\left[n - \sum_{i=1}^{k} P[i]\right],$$

where $G[k]$ are the corresponding gain coefficients and P the periods' values. Different stochastic models for jitter and shimmer have been proposed in the literature. In this work we assume, for a pulse train with a jitter $jitt\%$, a normal probability distribution for each period P:

$$pdf\,(P[k]) = \frac{1}{\sigma_P\sqrt{2\pi}} \exp\left(-\frac{(P[k]-P_0)^2}{2\sigma_P^2}\right),$$

where P_0 is the mean period and $\sigma_P = \frac{P_0\,jitt\,\%}{200}$. In order to avoid period approximation problems, a uniform randomized roundness function and a sampling frequency of 50 KHz have been used.

In a similar way, the gain coefficients distribution is given by:

$$pdf\,(G[k]) = \frac{1}{\sigma_G\sqrt{2\pi}} \exp\left(-\frac{(G[k]-1)^2}{2\sigma_G^2}\right).$$

Four hundred signals were synthesized, 100 corresponding to male and 100 to female, for each group of normal and pathological voices. For each situation, the model parameters were obtained from the statistics of real signals, adopting a fundamental frequency with a distribution $\mathcal{N}(144, 22.5)$ for male voices and $\mathcal{N}(245, 24.5)$ for female voices; a $\mathcal{N}(0.4, 0.1)$ jitter distribution for normal voices and $\mathcal{N}(5, 1)$ for pathological voices; and a shimmer with distribution $\mathcal{N}(1, 0.2)$ and $\mathcal{N}(8, 1)$ respectively.

2.2 Real Data

The implementation of the method proposed for estimation of F_0 was analyzed using a database of vowel signals from 710 persons of both genders [17]. It includes sustained phonation of the vowel /a/ of 53 healthy individuals and patients with a wide variety of voice disorders (organic, neurological, traumatic and psychogenic). The healthy voices belong to 21 males and 32 females, with mean ages 38.81 ± 8.49 and 34.16 ± 7.87 years, respectively. The set of 657 pathological voices contains samples of 169 male speakers, 238 female speakers and 247 without data about gender. The average ages are 49.80 ± 17.46 years for males and 46.83 ± 17.41 years for females. Some of the present disorders are adductor and abductor spasmodic dysphonia, A-P squeezing, cysts, erythema, gastric reflux,granulation tissue, hyperfunction, interaytenoid hyperplasia, keratosis / leukoplakia, paralysis, polypoid degeneration, scarring, ventricular compression, vocal fold edema, vocal fold edema, vocal fold polyp, vocal tremor, and others. In this database, the average fundamental frequency of normal voices is between 120.39 and 316.50 Hz.

For voice classification experiments a corpus of sustained vowels /a/ was used. The speech utterances from this corpus were registered in an anechoic room (global reverberation time $<$ 30 msec.). Each subject was requested to phonate the sustained vowels as steadily as possible toward an electrodynamic unidirectional microphone Shure SM58 at a distance of about 15 cm from the mouth. Each vowel had a duration of 1 to 5 sec. The data was digitized with a professional Turtle Beach Multisound FIJI sound card, at 44 KHz, 16 bits and no compression was used. Later, the data was low-pass filtered and down-sampled to 22 KHz. All the voices were classified by an experienced voice pathologist. It was considered a first set of 106 voices (half normal and half of diverse pathologies, randomly selected from a larger data base), here named Data Base DB1, and a second one of 14 normal voices, 13 of AdSD, and 6 of MTD, here named Data Base DB2.

Here it is important to point out that patients affected with AdSD may attempt to prevent their symptoms by increasing the tension in their laryngeal muscles in an effort to compensate their disease signs. The consequence is the appearance of additional physical disturbances similar to MTD along with AdSD. The over-riding symptoms of MTD can escalate over time making difficult to discern the underlying symptoms of AdSD [18].

2.3 EMD and EEMD

As it was stated in Sec. 1, EMD decomposes a signal $x(t)$ into a (usually) small number of IMFs. IMFs must satisfy two conditions: (i) the number

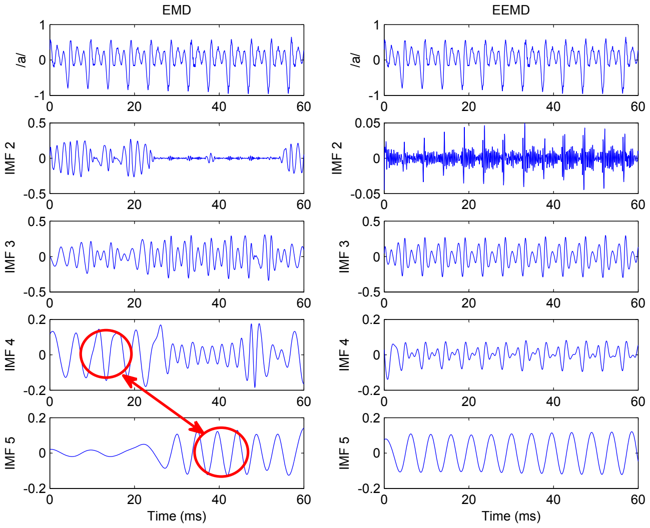

Fig. 1. A sustained real vowel /a/ corresponding to a normal subject, analyzed by EMD (left column) and EEMD (right column). The corresponding IMFs 2 to 5 are shown. In IMFs 4 and 5 corresponding to EMD two segments where "mode mixing" occurs, are marked with circles.

of extrema and the number of zero crossings must either be equal or differ at most by one; and (ii) at any point, the mean value of the upper and lower envelopes is zero.
Given a signal $x(t)$, the non-linear EMD algorithm, as proposed in [1], is described by the following algorithm:

1. find all extrema of $x(t)$,
2. interpolate between minima (maxima), obtaining the envelope $e_{min}(t)$ $(e_{max}(t))$,
3. compute the local mean $m(t) = (e_{min}(t) + e_{max}(t))/2$,
4. extract the IMF candidate $d(t) = x(t) - m(t)$,
5. check the properties of $d(t)$:
   – if $d(t)$ is not an IMF, replace $x(t)$ with $d(t)$ and go to step 1,
   – if $d(t)$ is an IMF, evaluate $r(t) = x(t) - d(t)$,
6. repeat the steps 1 to 5 by *sifting* the residual signal $r(t)$. The sifting process ends when the residue satisfies a predefined stopping criterion [4].

As already pointed out, one of the most significant EMD drawbacks for some applications is the so called mode mixing. It is illustrated in the left column of Fig. 1, where 60 ms of a sustained vowel /a/ are analyzed by EMD. The four IMFs with higher energy are shown. The appearance of oscillations of notoriously disparate scales in IMF 2 is clear. Another example can be seen in IMF 5, where oscillations are marked with circles. These oscillations are very similar to those marked on IMF 4.

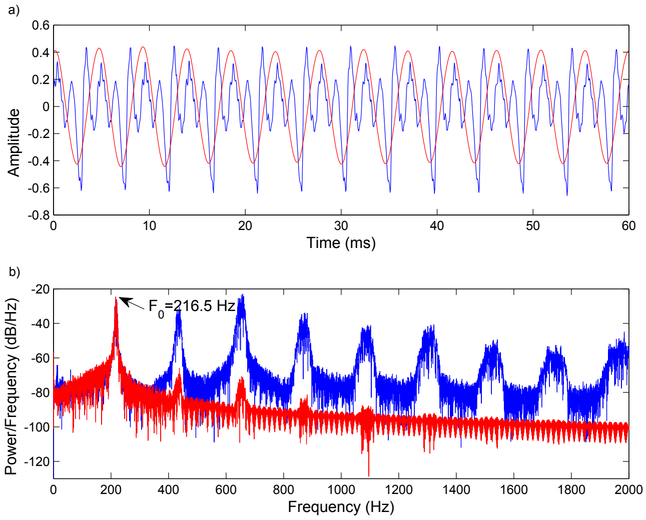

Fig. 2. a) A sustained real vowel /a/ corresponding to a normal subject (blue) and IMF 5, obtained by EEMD (red). b) PSD estimates of the sustained vowel /a/ (blue) and its EEMD based IMF 5 (red). The peak of the spectrum of the IMF 5 is marked as $F_0 = 216.5$ Hz.

EEMD[5], is an extension of the previously described EMD. It defines the true IMF components as the mean of certain ensemble of trials, each obtained by adding white noise of finite variance to the original signal. This method alleviates the mode mixing of the EMD algorithm [14]. An example of the EEMD abilities can be seen in the right column of Fig. 1. An ensemble size of $N_e = 5000$ was used, and the added white noise in each ensemble member had a standard deviation of $\epsilon = 0.2$. In general a few hundred of ensemble members provide good results [14]. The remaining noise, defined as the difference between the original signal and the sum of the IMFs obtained by EEMD, has a standard deviation $\epsilon_r = \epsilon/N_e$. For a complete discussion about the number of ensemble members and noise standard deviation, we refer to [14]. The IMFs 2 to 5 are shown in the right column of Fig. 1, below the sustained vowel /a/. The IMFs obtained by EEMD seem to be much more regular than the EMD version and, additionally, we can appreciate that in IMF 5 the oscillations capture the fundamental period of the sustained /a/.

This fact is remarked in Fig. 2.a, where the sustained vowel /a/ is pictured and the EEMD related IMF 5 is superimposed in a red line. In Fig. 2.b the power spectral densities (PSD) of /a/ and IMF 5 are plotted. The PSD of IMF 5 has a well defined peak in the frequency $F = 216.5$ Hz, which can be understood as a mean fundamental frequency. A visual inspection of the waveform (Fig. 2.a) allows the estimation of the funda-

[5] Matlab software available at `http://rcada.ncu.edu.tw/`.

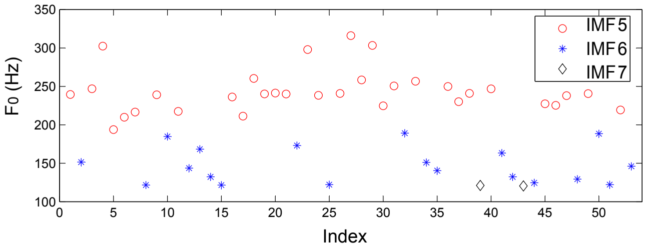

Fig. 3. F_0 average over the 53 analyzed sustained real vowels corresponding to normal subjects. Circles (red), stars (blue) and diamonds (black) indicate the signals in which F_0 was found in modes 5, 6 and 7 respectively.

mental frequency as close to 200 Hz, what is in agreement with the PSD of IMF 5.

3 Instantaneous Frequency Extraction

In this section we present and discuss the main ideas of the algorithm, based on EEMD, for the extraction of F_0.

Once the EEMD is computed, we want to identify the mode in which F_0 stands almost alone. With this in mind, a visual inspection of the decomposition of the normal voices in our database, allows to identify the candidate mode, as can be appreciated observing the second column in Fig. 1 and Fig. 2.a. Clearly F_0 is present in modes 3, 4 and 5. In the two first ones it is mixed with other components of the original signal, but it appears alone in the last one. This fact is reinforced by the sinusoidal like waveform of IMF5.

In our 53 samples of normal voices, F_0 was found in the IMFs 5, 6 and 7. Only in two cases it has been found in IMF 7, with average values of 120.394 Hz and 121.102 Hz, while in nineteen cases it was found in IMF 6, with average values in between 121.652 Hz and 189.295 Hz. In the remaining 32 voice, F_0 was found in IMF 5 with average values in between 193.934 Hz and 316.504 Hz.

Fig. 3 shows the average values of the instantaneous frequencies obtained from the modes identified by visual inspection in each normal voice in our database. In red circles are indicated those voices whose F_0 was identified at mode 5, while the blue stars and the black diamonds indicate those cases corresponding to an identification in modes 6 and 7 respectively. It can be appreciated that it exists a relationship between the F_0 average value and the mode in which F_0 has been identified. This is consistent with the results of Flandrin *et al.*[19]. They showed that, when applied to white noise, the EMD acts as an adaptive dyadic filter bank.

In order to obtain an automatic method to select the mode in which F_0 is hidden, we explore the discrimination abilities of the Shannon entropy in the present context. In the discrete case, it is defined as: $H(x) = -\sum_{i=1}^{M} p_i \log(p_i)$, with the understanding that $p\log(p) = 0$ if $p = 0$,

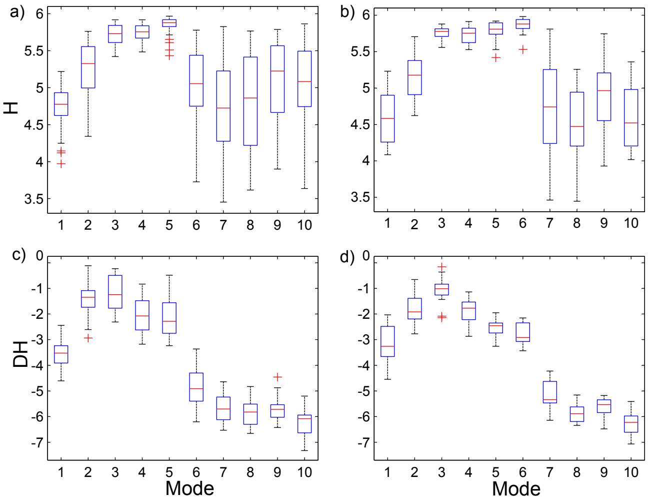

Fig. 4. a, b) Discrete entropies of modes 1 to 10 for the sustained real normal vowels /a/ at which F_0 was found in modes 5 and 6, respectively. c, d) Differential entropies of modes 1 to 10 for the sustained real normal vowels /a/ for which F_0 was found in modes 5 and 6, respectively.

where p_i is the probability that the signal x belongs to a considered interval and M is the partitions number [20].

Fig. 4.a displays the boxplots of the Shannon discrete entropy (H) corresponding to the ten first modes of the sustained normal vowels for which F_0 was found in mode 5.

In Fig. 4.b are shown those voices for which F_0 was found in mode 6. The histogram-based discrete entropy was estimated with 500 bins. It can be appreciated that the first mode has average entropy lower than for the other four or five modes (Figs.4.a and 4.b, respectively.) This is consistent with the fact that the first mode mainly contains high frequency noise: the one added to the original voices to perform the EEMD.

It can be observed that, for those voices for which F_0 is in IMF 5, the entropy has a jump after this mode, while a similar jump is observed in the 6th mode for those voices in which the fundamental frequency was found in IMF 6. There is however an overlap, which does not appear if we use an estimate of the differential entropy (DH) [21] instead of the discrete one. DH was estimated using a smoothing Gaussian kernel probability density estimation with 500 equally spaced points that cover the range of each IMF [21].

The results shown in Figs. 4.c and 4.d correspond to the differential entropy of those voices for which the fundamental frequency was found in modes 5 and 6. It should be noted here that in the case of the normal voices, the IMFs obtained through the EEMD have sinusoidal shapes and their probability density functions are also similar to a sinusoidal pdf. Therefore, if we remember that the differential entropy of a sinusoidal of amplitude A is given by $DH = \ln(\pi A/2)$, it would be reasonable to

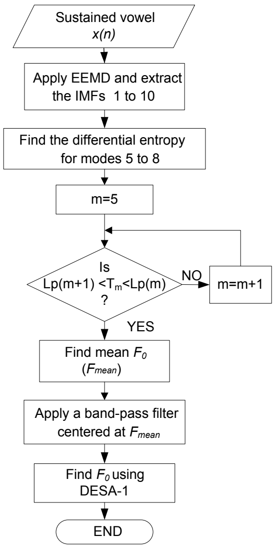
Sustained vowel
x(n)
Apply EEMD and extract the IMFs 1 to 10
Find the differential entropy for modes 5 to 8
m=5
Is Lp(m+1) <T$_m$<Lp(m) ?
NO
m=m+1
YES
Find mean F_0 (F_{mean})
Apply a band-pass filter centered at F_{mean}
Find F_0 using DESA-1
END

Fig. 5. Flow diagram corresponding to the full F_0 extraction method based on EEMD.

propose the logarithm of the power of the IMFs as an index to find the mode where F_0 is hidden. This idea will be addressed in future works. Taking into account these results, for modes $m = 5, 6$ and 7, we can propose thresholds T_5, T_6 and T_7 such that:$-3.365 < T_5 < -3.234$, $-4.224 < T_6 < -3.433$ and $-5.762 < T_7 < -4.172$. In this way, given a voice, if its DH corresponding to mode 5 is higher than T_5 and its DH corresponding to mode 6 is lower than T_5, it could be expected its F_0 would be hidden in mode 5. If this is not the case, the presence of a jump in between modes 6 and 7 should have to be tested using threshold T_6, and afterwards in between modes 7 and 8 by means of threshold T_7. This hypothesis should have to be tested on a larger data base, which is right now not available. This would allow setting more accurate thresholds. Once the mode where it is expected to find F_0 is selected, spurious components must be eliminated. For this task we adopt a type II Chebyshev bandpass filter, with a bandwith of 150 Hz and centered on the frequency corresponding to the maximum of the spectrum of the selected mode. This frequency is a good approximation of the average value of F_0, as shown in Fig. 2.b. At this point, an AM-FM separation algorithm must be applied. The DESA-1 [22] provides us better results than Hilbert-Transform based methods, as reported in [23]. The flow diagram corresponding to the full algorithm is displayed in Fig.5.

4 Pathological Voice Classification

In the last section the ensemble version of EMD was used in order to improve the possibility of finding the F_0 in a unique mode. Here, we will

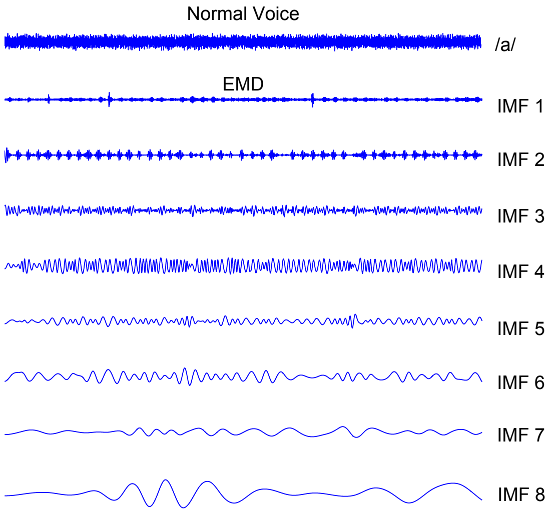

Fig. 6. Sustained real normal vowel /a/ in the first row and IMFs 1-8 of its EMD.

explore the non linear decomposition capabilities of EMD to produce some discriminative information useful for pathological voice classification.

In a previous work we selected eight standard acoustic parameters extracted from sustained vowels, including short-term perturbations of fundamental frequency and intensity (termed *jitter* and *shimmer*, respectively), and glottal noise measures [16]. These feature vectors were used to perform an automatic classification of normal and pathological voices, such as those here considered, improving the best reported results. In [27] we explored different dimensionality reduction techniques to perform the visualization and classification using similar feature vectors. Even if in this work we obtained good results for final vector of dimensions 2 and 3, it must be noticed that the physical meaning of each dimension was lost due to the transformations involved. Therefore in this work we explore a new EEMD based approach that could allow to reduce the dimensionality of the feature vector, retaining certain physical meaning of its components.

In our experiments, the EMD algorithm of sustained real vowels stopped at IMF 12 ± 1. As an example a sustained normal vowel /a/ and the first eight IMFs of its EMD are shown in Fig.6. Inspired by Fig. 7 we propose to consider for our classification and visualization purposes, a new feature vector in $\mathbb{R}^6$, which components are the maximum PSD of the IMFs 2-4 and the corresponding frequencies.

It must be emphasized that the EMD based algorithms act as an adaptive filter bank that is guided by the data [4], meaning that for signals with different frequency content, a given frequency can be found in different modes. This fact can reinforce the differences between the normal and pathological signals. In this way, the use of the proposed feature vectors

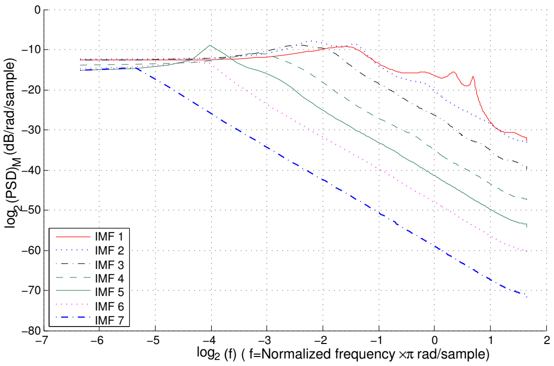

Fig. 7. Log-log power spectrum density, estimated with Burg algorithm, corresponding to each of the IMFs of a Spanish sustained real normal vowel /a/ displayed in the previous figure.

could allow to provide new information to improve the discrimination between this kind of data.

Using these new feature vectors, a K-nearest neighbors' classification rule was applied and a K-fold cross validation method, with 20 subsamples, was used in order to estimate the classifier performance.

5 Results

5.1 Instantaneous Frequency Extraction

Simulated Normal and Pathological Voices For illustration purposes, F_0 was extracted with the method proposed in Sec. 3 from both normal and pathological simulated sustained vowels /a/. The results are shown in Fig. 8. For comparison, two additional pitch extraction methods were applied to the same data and also shown in Fig. 8. The RAPT method (black) [24] was implemented using VOICEBOX[6] , while an autocorrelation-based method (blue) [25] was implemented using the PRAAT software[7]. The parameters involved in these two algorithms are the default ones. We can observe in Fig. 8 several evident errors in doubling or halving-period events both in RAPT and AC-based methods, specially for the pathological voice case (Fig. 8.b). On the contrary, the EEMD based method here proposed performs smoothly and without errors in both simulation conditions.

Real Normal and Pathological Voices As in the previous example, F_0 was extracted with the method proposed in Sec. 3 and the other

[6] VOICEBOX Matlab toolkit v. 1.18 (2008), `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[7] PRAAT v.5.0.32, `http://www.praat.org`.

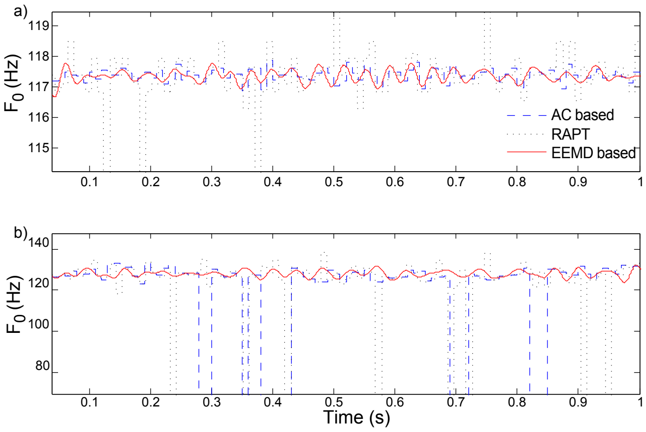

Fig. 8. F_0 of two simulated sustained vowels /a/ (generated as explained in Sec. 2.1) are analyzed: (a) normal simulated voice (b) pathological simulated voice. The results obtained by the autocorrelation based method (blue), RAPT (black) and the proposed EEMD based method (red) are shown.

two methods, from two sustained real normal vowels /a/. The results are presented in red in Figs. 9.a (EDC1NAL) and 9.b (JTH1NAL). Even if the results look similar, a careful inspection would reveal the above mentioned stair-case nature of the last two methods. This windowing artifact could be a problem for instantaneous frequency estimation.

The Pearson correlation coefficient between the mean F_0 of the 53 healthy sustained vowels /a/ reported in [17] and the averaged instantaneous frequency obtained by our method was $r = 0.999995$.

In Fig. 10 the F_0 corresponding to two pathological voices are presented. In Fig. 10.a the fundamental frequency of a sustained vowel /a/ from a patient suffering muscular tension dysphonia is analyzed with the proposed method. On the other hand, in Fig. 10.b a voice with adductor spasmodic dysphonia is studied. As in Fig. 9, the F_0 obtained with RAPT and auto-correlation based methods are also superposed in black and blue. Even if the autocorrelation based method had been reported to be the best pitch estimation technique for the analysis of pathological sustained vowel /a/ [26], it can be observed in that it fails several times (See Fig. 10). Also does RAPT algorithm, while the method here proposed, exhibits a much better behavior.

In a study with 35 disordered sustained vowels /a/ (15 from patients suffering muscular tension dysphonia and 20 suffering adductor spasmodic dysphonia) we have observed that, in the task of a correct F_0 extraction, while RAPT and auto-correlation based methods both fail in 22 voices (62.86 %), the here proposed algorithm reduced the number of failures to only 10 voices (28.57 %). The F_0 estimation was considered failed when at least one doubling or halving-period event, or a “spike-like” artifact appears. In the method here proposed, we have observed that these spike-like artifacts were coincident with pathological voice seg-

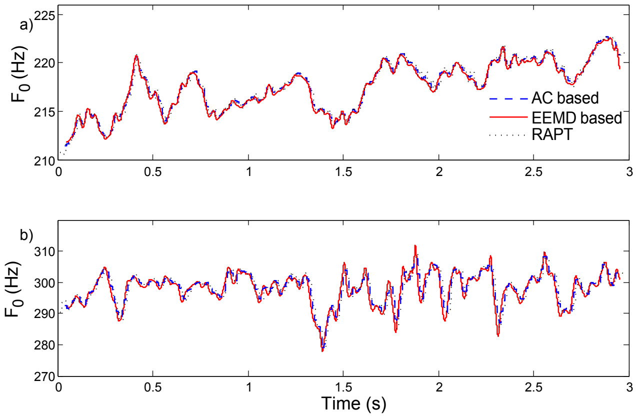

Fig. 9. F_0 of two healthy sustained vowels /a/ from database described in Sec. 2.2 are analyzed (a) EDC1NAL and (b) JTH1NAL. The results obtained by the autocorrelation based method (blue), RAPT (black) and the proposed EEMD based method (red) are shown.

ments of very low energy. In order to detect them and to prevent this kind of mistakes in the F_0 estimation, we consider that a voice-activity detection method could be applied as a pre-processing stage. However, the failures of the other two algorithms were more notorious. It is important to emphasize that the total length of the segments where the RAPT and autocorrelation-based methods fail, largely exceed the total length of all spike-like events related with the here proposed method. For this reason, if another quantifier is used in the algorithms comparison, as for example the percentage of signal length where the F_0 estimations are satisfactory, then the advantage of the EEMD based method would be more pronounced. These improvements will be addressed in future works.

5.2 Pathological Voice Classification

Simulated Normal and Pathological Voices In order to study the classification capability of the second new tool presented in Sec. 4, for each of the simulated voices we have selected as feature vectors' components the maximum PSD (log2) and the corresponding frequencies, of IMF i, $i = 2, 3, 4$.

With a simple and general-purpose classifier, a K-nearest neighbors' classification rule, the best performance was obtained using $K = 1$, reaching a 99% of correct classifications. In Table 1.a the obtained confusion matrix is presented. This result confirms that the IMFs' spectra provide relevant features that can be used as descriptors for the proposed classification task. The importance of this experiment is based on the fact that both normal and pathological synthetic voices have been simulated without added noise, and that the difference between them is only due

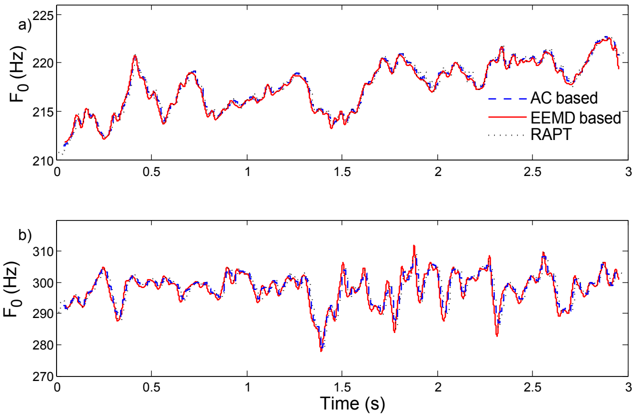

Fig. 10. F_0 of two pathological sustained vowels /a/ with: a) muscular tension dysphonia and b) spasmodic dysphonia. The results obtained by the autocorrelation based method (blue), RAPT (black) and the instantaneous F_0 estimated with the proposed EEMD based method (red) are shown.

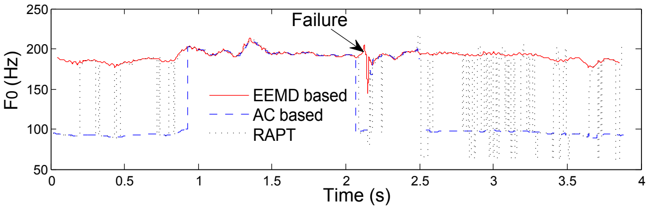

Fig. 11. F_0 of a pathological sustained vowel /a/. Autocorrelation method (blue), RAPT (black) and EEMD based method (red). Although the proposed method fails around $t = 2.1$ s, the other ones fails are more evident (period-doubling and period-halving errors).

to short-term perturbations of their fundamental frequency and intensity, as imposed in the model. Therefore, the proposed method is able to distinguish between normal and altered voices with very similar Fourier spectra. This a desirable property in the kind of pathologies we are dealing with.

Real Normal and Pathological Voices Following the same procedure as in the previous section, but with the real voices DB1, we obtained, with $K = 3$ a 93.40% of true positive classifications. In Table 1.b we present the corresponding confusion matrix, were we can appreciate that we obtained a 94.34% of correct classifications of the normal voices and a 92.45% in the pathological case. Taking into account that in Medicine, a pathological case is considered the positive one, these re-

Table 1. Confusion matrix

(a)	Simulated voices		
Class	Classifications		Correct
	Pathologic	Normal	Classifications
Pathologic	198	2	99%
Normal	2	198	99%
(b)	Real voices (DB1)		
Class	Classifications		Correct
	Pathologic	Normal	Classifications
Pathologic	49	4	92.45%
Normal	3	50	94.34%

sults indicate that the proposed method has a sensitivity of 0.925 and a specificity of 0.926.

In the case of discrimination between MTD and AdSD, we show some preliminary results that suggest that the new tools here presented could also be useful. Unfortunately the amount of data available at the present time is not enough to perform an appropriate statistical study from the point of view of signal analysis, even if from the medical point of view it is encouraging. Plotting for each voice the log2 values of the frequencies at which the maximum value of PSD is obtained for IMFs 2, 3, and 4, we can appreciate in Fig. 12.a) that it seems to be possible to separate AdSD from the normal and MTD. Plotting the maxima of the PSD (in log2), we see in Fig. 12.b) that it is possible to separate most of the MTD from the other pathology and the normal ones. Both plots collaborate to provide a possible separation in three classes. Therefore, these figures suggest that, if a larger set of data would be available, it could be possible to perform a first separation in two classes, class 1: ASD and class 2: normal and MTD, and them continue working on with class 2 to accomplish the final classification. We can appreciate that, for IMFs 2, 3, and 4, those voices in class 1 reach their maximum value of PSD at lower frequencies than the voice belonging to class 2. While MTD and normal voices could be separated just using the maxima of the PSD in IMF 3.

6 Discussion and Conclusions

In this work we have discussed some drawbacks and advantages of both the EMD and the EEMD and how both of these methods can be useful to extract relevant information from voiced signals. We have presented the abilities of EEMD for extracting the F_0 from sustained vowels /a/ in combination with an instantaneous frequency estimator (DESA-1) algorithm. Additionally, a technique for the automatic selection of the mode from which F_0 can be extracted was here proposed. The new method was successfully tested on normal and pathological sustained voices and compared with other algorithms. The EEMD based method has the ad-

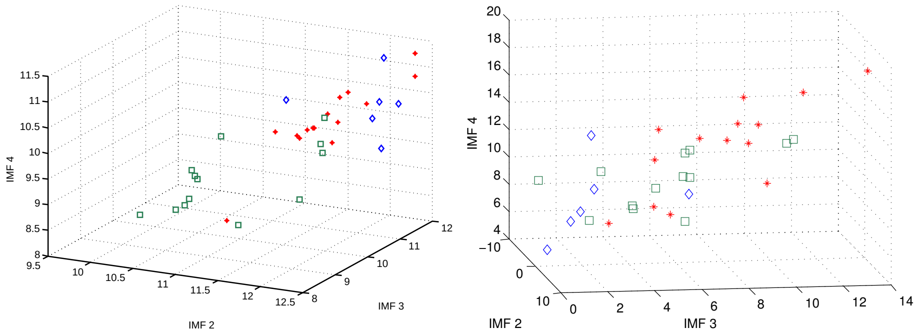

Fig. 12. (a) Frequency (log2) corresponding to the maximum Psd of three IMFs, of normal (stars) and pathological (diamonds – MTD – and squares – ASD) voices (DB2). (b) Maxima Psd (log2) corresponding to three IMFs of normal (stars) and pathological (diamonds – MTD – and squares – ASD) voices (DB2).

vantage to be parameters free, what is an interesting property for non-computational expert operators. As a drawback, the proposed method inherits the high computational cost of the EEMD algorithm. However, its utility in research and clinical applications without the need of on-line F_0 estimation is clear. These preliminary results suggest important advantages of the method here proposed and encourage us to continue the research on these ideas. Although very promising, all the conclusions here presented need to be statistically tested on a larger database. An extension to spontaneous speech and noisy signals will be addressed in future works.

We have also introduced a new method to discriminate between normal and pathological speech signals based on the spectral analysis of the IMFs obtained by means of EMD. We have applied this new tool to the analysis of speech signals corresponding to sustained vowels of different data sets: real and simulated voices. Inspired by the analysis of real data, we have performed an automatic classification of simulated voices (normal and pathologic), with a high accuracy rate (99.00%). In the case of discrimination between normal and pathological real voices we have obtained a performance of (93.40%).

The synthetic stimuli are generated by very simple LPC-synthesis excited by an impulse train. The real excitation spectrum of a voice is more complicated and would probably be a more difficult and realistic test to the proposed methods. This is confirmed by the fact that with the synthetic stimuli, the classifier has an accuracy rate of 99 % compared to 93 % with real voices. In future works we propose to use a more realistic glottal flow waveform as excitation.

We consider that it could be possible to overcome the best reported value by refining the proposed method. These preliminary results strongly suggest that spectral tools based on EMD are useful for the discrimination between normal and pathological voices. Moreover, they suggest that it could be possible to develop an automatic tool for differentiation between pathologies. Future works of this group include the application of these results to a wider data base of real signals, in continue collaboration with voice pathologists, and the analysis and discussion of other classification techniques.

Acknowledgments

This works was supported by UNER (Argentina), ANPCyT (Argentina) and CONICET (Argentina). The authors would like to thank Dr. M. C. Jackson-Menaldi of Lakeshore Prof. Voice Center, Lakeshore Ear, Nose and Throat Center, St. Clair Shores (USA) and Depart. of Otolaryng., School of Med., Wayne U., Detroit (USA), for her valuable suggestions. The authors also thank Kay Elemetrics Corp.

References

1. Huang, N., Shen, Z., Long, S., Wu, M., Shih, H., Zheng, Q., Yen, N., Tung, C., Liu, H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc.: Math., Phys. and Eng. Sciences **454** (1998) 903–995
2. Huang, N.E., Shen, S.S.P., eds.: Hilbert-Huang transform and its applications. Volume 5 of Interdisciplin. Math. Sc. World Sci.(2005)
3. Schlotthauer, G., Torres, M.E.: Descomposición modal empírica: análisis y disminución de ruido en señales biológicas. In: Proc. XV Congreso Argentino de Bioingeniería SABI 2005, Paraná, E.R. Argentina (2005) File:101PS.pdf ISBN 950-698-155-8.
4. Rilling, G., Flandrin, P., Gonçalvès, P.: On empirical mode decomposition and its algorithms. In: Proc IEEE-EURASIP Workshop NSIP-03, Grado, Italia (2003)
5. Rilling, G., Flandrin, P.: On the influence of sampling on the empirical mode decomposition. In: IEEE Int. Conf. On Acoust., Speech and Signal Proc. ICASSP-06. Volume III., Toulouse (2006) 444–447
6. Dimitriadis, D., Maragos, P.: Continuous energy demodulation methods and application to speech analysis. Speech Commun. **48**(7) (2006) 819–837
7. Schlotthauer, G., Torres, M.E., Rufiner, H.: A new algorithm for instantaneous F_0 speech extraction based on ensemble empirical mode decomposition. In: Proc. of 17th Eur. Sign. Proces. Conf. 2009, Glasgow, UK (2009) 2347–2351
8. Schlotthauer, G., Torres, M.E., Rufiner, H.: Voice fundamental frequency extraction algorithm based on ensemble empirical mode decomposition and entropies. In: Proc. of 11th Int. Congr. of the IFMBE 2009, Munich (2009) 984–987
9. Torres, M.E., Schlotthauer, G., Rufiner, H.L., Jackson-Menaldi, M.C.: Empirical mode decomposition. spectral properties in normal and pathological voices. In: Proc. of the 4th Eur. Conf. of the Inter. Fed. for Med. and Biol. Eng. (2009) 252–255
10. Hess, W.: Pitch and Voicing Determination of Speech with an Extension Toward Music Signals. In: Springer Handbook of Speech Proc. Springer (2008) 181–208

11. Schlotthauer, G., Torres, M.E., Jackson-Menaldi, M.C.: A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. J. of Voice (2009) In press.
12. Huang, H., Pan, J.: Speech pitch determination based on Hilbert-Huang transform. Signal Process. **86**(4) (2006) 792–803
13. Weiping, H., Xiuxin, W., Yaling, L., Minghui, D.: A Novel Pitch Period Detection Algorithm Bases on HHT with Application to Normal and Pathological Voice. In: IEEE-EMBS 2005. 27th Annual Intern. Conf. of the. (2005) 4541–4544
14. Wu, Z., Huang, N.E.: Ensemble empirical mode decomposition: A noise-assisted data analysis method. Adv. in Adapt. Data Anal. **1**(1) (2009) 1–41
15. Verdolini, K., Rosen, C.A., Branski, R.C., Andrews, M.L.: Classification Manual for Voice Disorders-I. 1 Edn. Lawrence Erlbaum Assoc. (2006)
16. Schlotthauer, G., Torres, M.E., Jackson-Menaldi, C.: Automatic diagnosis of pathological voices. WSEAS Trans. on Signal Proc. **2** (2006) 1260–1267 (And references therein).
17. Corp., K.E.: Disordered voice database 1.03. Mass. Eye and Ear Infirmary, Voice and Speech Lab, Boston. (1994)
18. Jackson-Menaldi, M.C.: La voz patológica. Ed. Médica Panamericana, Buenos Aires (2002)
19. Flandrin, P., Rilling, G., Gonçalvès, P.: Empirical mode decomposition as a filter bank. Signal Proc. Lett., IEEE **11**(2) (Feb. 2004) 112–114
20. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27** (1948) 379–423, 623–656
21. Papoulis, A.: Probability, Random Variables and Stochastic Processes. 3rd edn. McGraw-Hill Companies (1991)
22. Maragos, P., Kaiser, J., Quatieri, T.: Energy separation in signal modulations with application to speech analysis. Signal Proc., IEEE Trans. on **41**(10) (Oct 1993) 3024–3051
23. Diaz, M., Esteller, R.: Comparison of the non linear energy operator and the hilbert transform in the estimation of the instantaneous amplitude and frequency. Latin Am. Trans., IEEE (Revista IEEE America Latina) **5**(1) (2007) 1–8
24. Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Speech Coding and Synth. Elsevier Science (1995) 121–173
25. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. of the Inst. of Phonetic Sci. Volume 17. (1993) 97–110
26. Jang, S., Choi, S., Kim, H., Choi, H., Yoon, Y.: Evaluation of performance of several established pitch detection algorithms in pathological voices. Proc. 29th Annual Intern. Conf. of the IEEE Eng. in Med. and Biol. Soc. **2007** (2007) 620–623 PMID: 18002032.
27. Goddard, J., Schlotthauer, G., Torres, M.E., Rufiner, H.L.: Dimensionality reduction for visualization of normal and pathological speech data. Biomed. Sig. Proc. and Control **4** (2009) 194–201.