A NEW ALGORITHM FOR INSTANTANEOUS F_0 SPEECH EXTRACTION BASED ON ENSEMBLE EMPIRICAL MODE DECOMPOSITION

Gastón Schlotthauer, María Eugenia Torres and Hugo L. Rufiner

Laboratorio de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Universidad Nacional de Entre Ríos
CC 47 - Suc 3, 3100, Paraná, ARGENTINA
phone: + (54) 3434975100, fax: + (54) 3434975077, email: gschlott@bioingenieria.edu.ar; metorres@santafe-conicet.gov.ar

ABSTRACT

In this work, a new instantaneous fundamental frequency extraction method is presented, with the attention especially focused on its robustness for pathological voices processing. It is based on the Ensemble Empirical Mode Decomposition (EEMD) algorithm, which is a completely data-driven method for signal decomposition into a sum of AM - FM components, called Intrinsic Mode Functions (IMFs) or modes. Our results show that the speech fundamental frequency can be captured in a single IMF. We also propose an algorithm for selecting the mode where the fundamental frequency can be found, based on the logarithm of the power of the IMFs. The instantaneous frequency is then extracted by means of well-known techniques. The behaviour of the proposed method is compared with other two ones (Robust Algorithm for Pitch Tracking -RAPT- and auto-correlation based algorithms), both in normal and pathological sustained vowels.

1. INTRODUCTION

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 is $F_0 = 1/T_0$ [1]. Even if F_0 is useful for a wide range of applications, its reliable estimation is still considered one of the most difficult tasks. This is especially true in the presence of noise or in pathological voices [1]. In speech, F_0 variations contribute to prosody, and in tonal languages, as in Mandarin Chinese, they also help to distinguish segmental categories. Attempts to use F_0 in speech recognition systems have met with mixed success. In part, this may be a consequence of the lack of reliable estimation algorithms. Other current applications are related with speaker recognition, speech based emotion classification, voice morphing, singing and pathological voice processing.

A reliable and accurate estimation of F_0 is essential for a correct frequency perturbation analysis (also known as *jitter*). In the case of sustained vowel waveforms, this analysis is a standard procedure in the clinical evaluation of disordered voices, in assessing the severity of pathological voices, and in monitoring patient progress during treatment [2].

Conventional F_0 extraction algorithms are based on windowed segments, usually providing stair-case time series [1]. However, in pathological voice analysis it is desirable to have a smooth and accurate F_0 time series. Additionally, these methods assume that speech is produced by a linear system and that speech signals are locally stationary, two inappropriate oversimplifications in the case of pathological voices.

EMD has been recently proposed by Huang [3] for adaptively decomposing nonlinear and non stationary signals into a sum of *well-behaved* AM - FM components, called Intrinsic Mode Functions. This new technique has received the attention of the scientific community, both in its understanding and application. The method consists in a local and fully data-driven splitting of a (possibly non-stationary) signal in fast and slow oscillations. While in [4] six fixed band pass filters are used in order to obtain an AM FM model of speech, the EMD adaptively decomposes the speech signal into a sum of AM - FM components.

A few EMD based algorithms have been proposed for F_0 extraction [5; 6]. However, they suffer the well-known "mode mixing" problem and they use a set of post-processing rules with the intention of alleviate it [5].

The mode mixing is perhaps the major drawback of the original EMD. This effect is defined as a single IMF either consisting of signals of widely disparate scales (energies), or a signal of a similar scale residing in different IMF components [7]. Wu and Huang [7] proposed a modification to the EMD algorithm. This new method, called Ensemble Empirical Mode Decomposition (EEMD), largely alleviates the mode mixing effect.

In this paper we present a new method based on EEMD which is able to extract the instantaneous F_0 in normal and pathological sustained vowels.

2. MATERIALS AND METHODS

2.1 Database

The database [8] developed by Massachusetts Eye and Ear Infirmary (MEEI) was used as test database. It contains voice samples of 710 subjects. Included are sustained phonation speech samples of the vowel /a/ from patients with a wide variety of organic, neuralgic, traumatic, and psychogenic voice disorders, as well as 53 normal subjects. There are both male and female cases in each group of pathologies. In the case of normal voices, the lowest mean fundamental frequency is 120.39 Hz and the highest mean fundamental frequency is 316.50 Hz. All signals were downsampled to 22050 Hz.

2.2 Ensemble Empirical Mode Decomposition

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 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 [3], is described by the following algorithm:

1. find all extrema of $x(t)$,

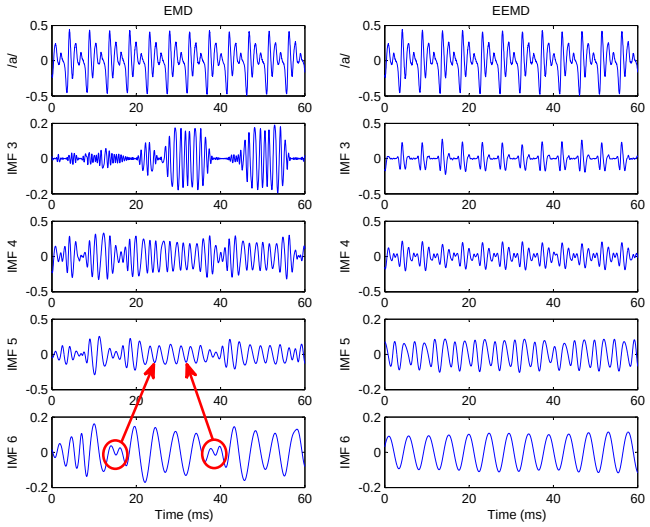

Figure 1: A sustained vowel /a/ analyzed by EMD (left column) and EEMD (right column). The corresponding IMFs 4 to 6 are shown. In IMF 6 corresponding to EMD two segments where "mode mixing" occurs are marked with circles.

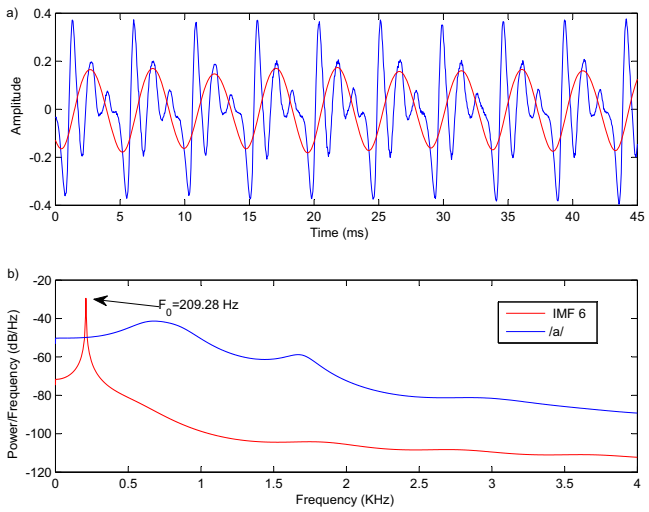

Figure 2: a) Sustained vowel /a/ (blue) and IMF 6, obtained by EEMD (red). b) PSD estimates of the sustained vowel /a/ (blue) and its EEMD based IMF 6 (red). The peak of the spectrum of the IMF 6 is marked as $F_0 = 209.28$ Hz.

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 [9].

As already pointed out, one of the most significant EMD drawbacks is the so called mode mixing. It is illustrated in the left column of Fig. 1, where a frame of 60 ms length of a sustained vowel /a/ is analysed by EMD. The four IMFs with higher energy are shown. The appearance of oscillations of dramatically disparate scales in IMF 3 is clear. Another example can be seen in IMF 6, where two oscillations are marked with circles. These oscillations are very similar to those on IMF 5.

EEMD[1], is an extension of the previously described EMD. It defines the true IMF components as the mean of certain ensemble of trials, each one obtained by adding white noise of finite variance to the original signal. This method provides a major improvement on the EMD algorithm, alleviating the mode mixing [7]. 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 Gaussian noise in each ensemble member had a standard deviation of $\varepsilon = 0.2$. In general a few hundred of ensemble members provide good results [7]. The remaining noise, defined as the difference between the original signal and the sum of the IMFs obtained by EEMD, has a standard deviation $\varepsilon_r = \varepsilon/N_e$. For a complete discussion about the number of ensemble members and noise standard deviation, we refer to [7]. The IMFs 3 to 6 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 6 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 6 is superimposed in a red line. In Fig. 2.b the power spectral densities (PSD) of vowel /a/ and its IMF 6 are plotted. The PSD of IMF 6 have a well defined peak in the frequency $F = 209.28$ Hz, which can be understood as a mean fundamental frequency. A visual inspection of the sonogram (Fig. 2.a) allows estimating that the fundamental frequency is near 200 Hz, what is consistent with the PSD of IMF 6.

2.3 Discrete Energy Separation Algorithm (DESA-1)

Once the IMFs are obtained, a method must be selected in order to separate the instantaneous amplitude and frequency. Usually Hilbert transform (HT) based techniques are used. However the Discrete Energy Separation Algorithm (DESA-1) overcomes the HT methods when actual signals are considered [10].

Let $d^m(n)$ be a sampled version of a continuous IMF, with $n = 1,\ldots,N$, for $m = 1,\ldots,M_x$, where M_x indicates the number of modes in which $x(t)$ is decomposed.

Then, we can define the discrete Teager energy operator by [11] $\Psi[d^m(n)] = (d^m(n))^2 - d^m(n-1)d^m(n+1)$, for $n = 2,\ldots,N-1$.

If $d^m(n)$ is a discrete time cosine with constant amplitude A and frequency ω, $d^m(n) = A\cos(\Omega n + \theta)$, with $\Omega = \omega T$ and T the sampling period, then:

$$\Psi[d^m(n)] \quad = \quad A^2\omega^2\left(\frac{\sin\Omega}{\Omega}\right)^2.$$

Based on these relations, we apply the DESA-1 for AM-FM separation [11]. It estimates the instantaneous frequency

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

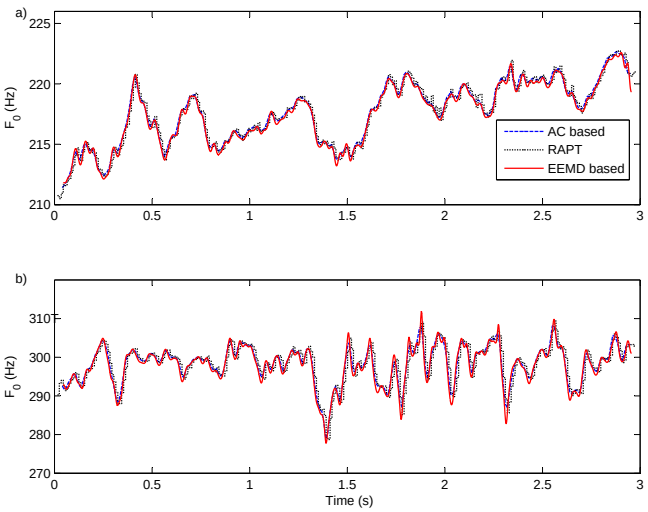

Figure 3: F_0 of two healthy sustained vowels /a/ from database described in Sec. 2.1 are analyzed (a) EDC1NAL and (b) JTH1NAL. The results obtained by the autocorrelation based method (black), RAPT (blue) and the instantaneous F_0 estimated with the proposed EEMD based method (red) are shown.

$\Omega(n)$ and the instantaneous envelope $a(n)$ by:

$$\Omega(n) = \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[d^m(n)]}\right),$$

$$|a(n)| = \sqrt{\frac{\Psi[d^m(n)]}{1 - \left(1 - \frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[d^m(n)]}\right)^2}},$$

where $y(n) = d^m(n) - d^m(n-1)$ for $n = 2, \ldots, N$.

3. RESULTS AND DISCUSSION

Visual inspection of each of the IMFs obtained by EEMD, for each of the normal voices (see Sec. 2.1), was carried out in order to find the mode that includes the instantaneous frequency (see Fig. 2.a). For illustration purposes, F_0 was extracted with the method proposed in the previous section from two sustained vowels /a/. These results are presented in red in Figs. 3.a (EDC1NAL) and 3.b (JTH1NAL). For comparison, two additional pitch extraction methods were applied to the same normal voice records and also shown in Fig. 3. The RAPT method (black) [12] was implemented using the VOICEBOX Toolkit [2], while an autocorrelation-based method (blue) [13] was implemented using the PRAAT software [3]. The parameters involved in these two algorithms are the default ones. It can be observed that the results were similar, although a carefully inspection reveals 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 [8] and the

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

[3]PRAAT v. 5.0.32(2008), available at `http://www.praat.org`.

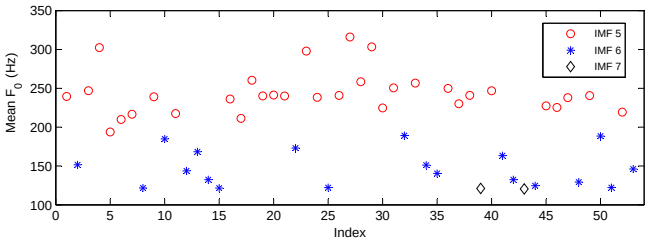

Figure 4: Mean F_0 of the 53 analyzed normal sustained vowels /a/. Red circles, blue stars, and black diamonds indicate the records where F_0 were founded in IMF 5, 6, and 7 respectively.

Table 1: Relationship between mean F_0 and the IMF.

Average F_0 (min - max) Hz	IMF	Occurrences in DB
120.394 - 121.102	7	2
121.652 - 189.295	6	19
193.934 - 316.504	5	32

averaged instantaneous frequency obtained by our method was $r = 0.999995$.

In the case of the 53 normal sustained vowels here considered, we obtained that the fundamental frequency was embedded in the fifth, sixth or seventh IMF. In Fig. 4 we show the average values of the instantaneous F_0, estimated using the proposed method, for indexes from 1 to 53, corresponding to each one of the normal voices on the database. Voices where the F_0 was found in IMF 5 were represented with red circles, while with blue stars and black diamonds were represented those voices where the F_0 were respectively in the IMFs 6 and 7. An interesting matching on the mean values range can be observed in each case.

In agreement with the studies of Flandrin *et al.*, which showed that the EMD is effectively an adaptive dyadic filter bank when applied to white noise [14], the IMFs containing the F_0 depends on its mean value.

This relationship is presented in Table 1, where the results of analyzing the normal sustained vowels /a/ from the Kay Elemetrics database [8] are presented. In two occasions the F_0 was found in IMF 7, with averages 120.394 Hz and 121.102 Hz. The F_0 was encountered in IMF 6 nineteen times, with averages between 121.652 Hz and 189.295 Hz. Finally, the F_0 was in IMF 5 in the 32 remaining voices, averaging between 193.934 Hz and 316.504 Hz.

In the case of a previously unobserved signal and without information about the mean F_0, a method is necessary in order to decide what is the IMF containing the F_0. In Fig. 5.a and Fig. 5.b, two boxplots graphics of the logarithm of the IMFs powers are presented. The boxplot shown in Fig. 5.a was estimated with the 32 sustained vowels where F_0 is in IMF 5, while Fig. 5.b was estimated using the 19 sustained vowels where F_0 is in IMF 6. Indeed, a clear step exists between the logarithm of the power of the mode where the F_0 is present and the next one. This finding can be used as an indicator to point out at which mode the F_0 should be looked for. Based on these results, for each mode = 5, 6, and 7, we can propose the thresholds T_5, T_6, and T_7 for normal voices as the followings: $-9.315 < T_5 < -9.093$, $-11.200 < T_6 < -9.509$, and $-10.970 < T_7 < -9.186$. In this way, if the logarithm of the power of IMF 5 is higher

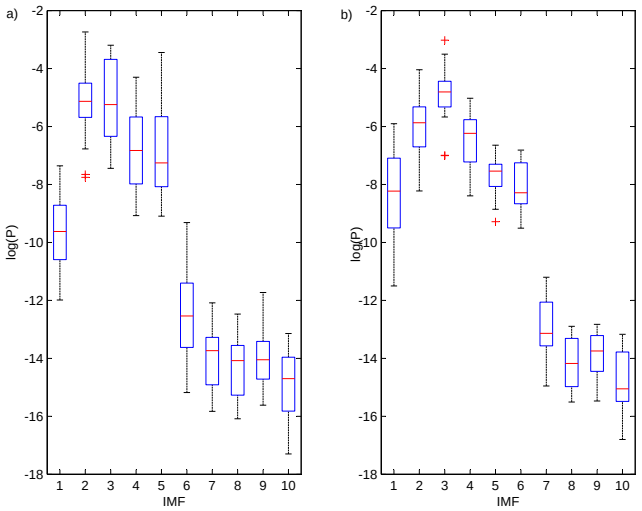

Figure 5: Boxplot of the logarithm of the powers of IMFs 1 to 10 of normal sustained vowels /a/ in which F_0 is present in IMF 5 (a) and in IMF 6 (b).

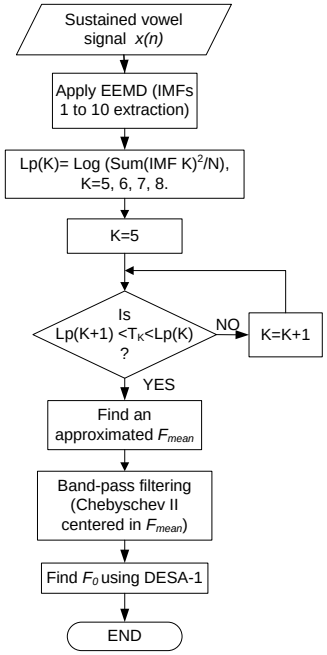

Figure 6: EEMD based F_0 extraction algorithm.

than T_5 while the logarithm of the power of IMF 6 is lower than T_5, then F_0 is expected to be in IMF 5. Otherwise, we test the existence of a step between IMF 6 and 7 using T_6, and next between IMFs 7 and 8 using T_7. In order to confirm this hypothesis and to obtain optimum thresholds, this study should have to be carried on with a larger data set.

Once we found the mode where F_0 is included, and with the purpose of eliminating spurious frequency components, a band-pass Chebyshev Type II filter is applied on it. This filter is centered in the frequency corresponding to the maximum of the PSD of the selected IMF. As in the example shown in Fig. 2.b, this frequency is a good approximation to the mean of the F_0. We selected a 150 Hz filter bandwidth. The complete algorithm here proposed is in this way obtained. In Fig. 6 the corresponding flow diagram is presented.

In voice pathology assessment, several parameters extracted from pitch estimation are commonly used [2]. Then, it is very important to have a good and reliable F_0 estimation. Unfortunately, there is no F_0 extraction method which operates consistently for pathological voices. This is due to the more serious and complex irregularities of vocal folds vibration in pathological voices than in normal. Many difficulties arise when estimating F_0, especially when pathological voices are analyzed, including period-doubling and period-halving.

In Fig. 7 the F_0 corresponding to two pathological voices are presented. In Fig. 7.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. 7.b a voice with adductor spasmodic dysphonia is studied. As in Fig. 3, the F_0 obtained with RAPT and auto-correlation based methods are also superposed in blue and black. Even if the autocorrelation based method had been reported to be the best pitch estimation technique for the analysis of pathological sustained vowel /a/ [15], it can be observed that it fails several times (see Fig. 7). Also does RAPT algorithm, while the method here proposed, exhibits the best behaviour.

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 fail both in 22 voices (62.86%), the here proposed algorithm reduced the number of fails to only 10 voices (28.57%). It must be cleared that in this study, we have considered as a failure any "spike-like" event as the one displayed in Fig. 8. In this figure, the F_0 of a sustained vowel /a/ from a patient suffering muscular tension dysphonia was extracted using the EEMD based, auto-correlation and RAPT methods. These kind of "spike-like" events are present in voice segments where the energy is very low. The instantaneous amplitude could be estimated as a pre-processing in order to detect these voice segments and to prevent this kind of mistakes in the F_0 estimation. These improvements will be addressed in future works. However, the failures of the other two algorithms are more evident. To address period-doubling problems, other performance quantifier score, like the period of time where the F_0 estimations are satisfactory, should be applied. Then the advantage of our EEMD based method would be much more pronounced.

4. CONCLUSIONS AND FUTURE WORK

In this work 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 advantage to be parameters free, what is an interesting property for non-computational expert operators. As a drawback, the proposed method has a high computational cost. However, we are mainly concerned with its utility in research and clinical applications without the need of on-line F_0 estimation.

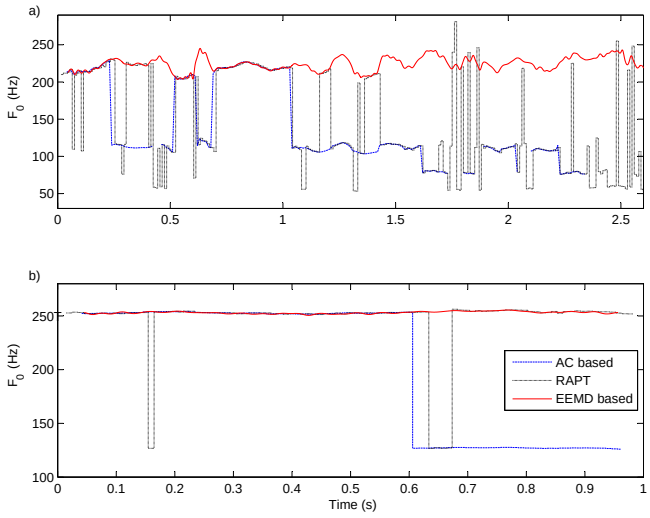

Figure 7: F_0 of two pathological sustained vowels /a/ with a) muscular tension dysphonia and b) spasmodic dysphonia. The results obtained by the autocorrelation based method (black), RAPT (blue) and the instantaneous F_0 estimated with the proposed EEMD based method (red) are shown.

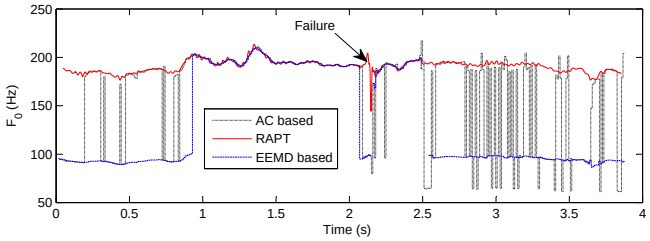

Figure 8: F_0 of a pathological sustained vowel /a/. The results obtained by the autocorrelation based method (black), RAPT (blue) and the instantaneous F_0 estimated with the proposed EEMD based method (red) are shown. Even if a failure in the EEMD based method can be observed around $t = 2.1$ s, the two other methods fail in a more obvious way, evidenced by the period-doubling phenomena.

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.

Acknowledgments

This works was supported by PID-UNER 6107-2 and PID-UNER 6111-2 (Universidad Nacional de Entre Ríos, Argentina) and by PAE 37122 and PAE-PICT-2007-00052 (Universidad Nacional del Litoral, Universidad Nacional de Entre Ríos, and National Council of Scientific and Technical Research -CONICET-, Argentina). The authors would like to thank Dr. María Cristina Jackson-Menaldi of Lakeshore Professional Voice Center of the Lakeshore Ear, Nose and Throat Center, St. Clair Shores (USA) and Depart. of Otolaryngology, School of Medicine, Wayne State University, Detroit (USA), for her valuable suggestions.

References

[1] W.J. Hess, "Pitch and voicing determination of speech with an extension toward music signals," in J. Benesty, M.M. Sondhi, and Y. Huang (Eds.): *Springer handbook of speech processing*, Springer-Verlag, 2008.

[2] G. Schlotthauer, M.E. Torres, M.C. Jackson Menaldi, "A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification," *J. Voice*, in press, 2009.

[3] N.E. Huang, Z. Shen, S. R. Long, M. L. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung and H. H. Liu, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 454, pp. 903–995, 1998.

[4] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech Communication*, vol. 48, pp. 819-837, 2006.

[5] H. Huang and J. Pan, "Speech pitch determination based on Hilbert-Huang transform," *Signal Processing*, vol. 86, pp. 792-803, 2006.

[6] H. Weiping, W. Xiuxin, L. Yaling, and D. Minghui, "A Novel Pitch Period Detection Algorithm Bases on HHT with Application to Normal and Pathological Voice," in *Proc. 2005 IEEE Eng. Med. and Biol. 27th Annual Conf.*, Shanghai, China, September 1-4. 2005, pp. 4541–4544.

[7] Z. Wu and N.E. Huang, "Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, pp. 1-41, 2009.

[8] Kay Elemetrics Corp.,"Disordered Voice Database ver. 1.03," Massachusetts Eye and Ear Infirmary, Voice and Speech Lab, Boston, 1994.

[9] G. Rilling, P. Flandrin, and P. Gonçalvès, "On empirical mode decomposition and its algorithms," in *Proc. 2003 IEEE-EURASIP Workshop NSIP-03*, Grado, Italy, Jun. 8-11. 2003.

[10] M. Díaz and R. Esteller, "Comparison of the non linear energy operator and the Hilbert transform in the estimation of the instantaneous amplitude and frequency," *IEEE Latin America Transactions*, vol. 5, pp. 1-8, 2007.

[11] P. Maragos, J.F.Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans Signal Proc*, vol. 41, pp. 3024-3051, 1993.

[12] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding & Synthesis*, W.B. Kleijn and K.K. Paliwal (Eds.), 1995.

[13] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of the Institude of Phonetics Sciences*, vol. 17, pp. 97-110, 1993.

[14] P. Flandrin, G. Rilling, and P. Gonçalvès, "Empirical mode decomposition as a filter bank," *IEEE Signal Processing Lett.*, vol. 11, pp. 112–114, 2004.

[15] S.-J. Jang, S.-H. Choi, H.-M. Kim, H.-S. Choi, and Y.-R. Yoon, "Evaluation of Performance of Several Established Pitch Detection Algorithms in Pathological Voices," in *Proc. of the 29th Annual Int. Conf. IEEE-EMBS*, Lyon, France, Aug. 23-26. 2007, pp. 620–623.