Voice Fundamental Frequency Extraction Algorithm Based on Ensemble Empirical Mode Decomposition and Entropies

G. Schlotthauer[1,2], M.E. Torres[1,2,3] and H. L. Rufiner[1,2,3]

[1] Laboratorio de Señales y Dinámicas no Lineales, Fac. de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde (E. R.), Argentina
[2] National Council of Scientific and Technical Research (CONICET), Argentina.
[3] Lab.de Señales e INteligencia Computacional (SINC), Fac. de Ingeniería y Cs Hídricas, Univ. Nac. del Litoral, Santa Fe, Argentina

Abstract— **A new algorithm for pitch extraction based on the Ensemble Empirical Mode Decomposition (EEMD) is presented. Applications to normal and pathological voices are considered. EEMD is a completely data-driven method for signal decomposition into a sum of AM - FM components, called Intrinsic Mode Functions (IMFs) or modes, which can be written as $A(t)\cos(\varphi(t))$. The voice fundamental frequency (F_0) can be captured in a single IMF, allowing its extraction by means of well known AM-FM separating techniques. An entropy based selection algorithm is here proposed, in order to determine the mode that holds the fundamental frequency. The behavior of the proposed method is compared with other two ones, both in normal and pathological sustained vowels.**

Keywords— **Ensemble empirical mode decomposition, fundamental frequency, pathological voice, entropy.**

I. 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 or pitch 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. 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 [2]. For this application, a reliable and accurate estimation of F_0 is essential. Conventional F_0 extraction algorithms are based on windowed segments, usually providing stair case time series [1]. These methods assume that speech is produced by a linear system and that speech signals are locally stationary. However, in the above mentioned applications it is desirable to have a smooth and accurate F_0 time series.

In voice pathology assessment, several parameters extracted from pitch estimation are commonly used. 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.

Empirical Mode Decomposition (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 (IMFs). 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 "mode mixing" problem. Wu and Huang [7] proposed a modification to the EMD algorithm, called Ensemble EMD (EEMD), which largely alleviates this effect. Here we present a new method based on EEMD which is able to extract the F_0 in normal and pathological sustained vowels.

II. MATERIALS AND METHODS

A. Database

As test database we use [8]. It includes 710 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. For these normal voices, the average F_0 is in the range from 120.39 Hz to 316.50 Hz. All signals were downsampled to 22050 Hz.

B. Ensemble Empirical Mode Decomposition

As already stated, EMD decomposes a signal *x(t)* into a (usually) small number of IMFs. They 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;

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. Schlotthauer, M. E. Torres & H. L. Rufiner; "Voice Fundamental Frequency Extraction Algorithm Based on Ensemble Empirical Mode Decomposition and Entropies"
Proc. of the World Congress on Medical Physics and Biomedical Engineering, sep, 2009.

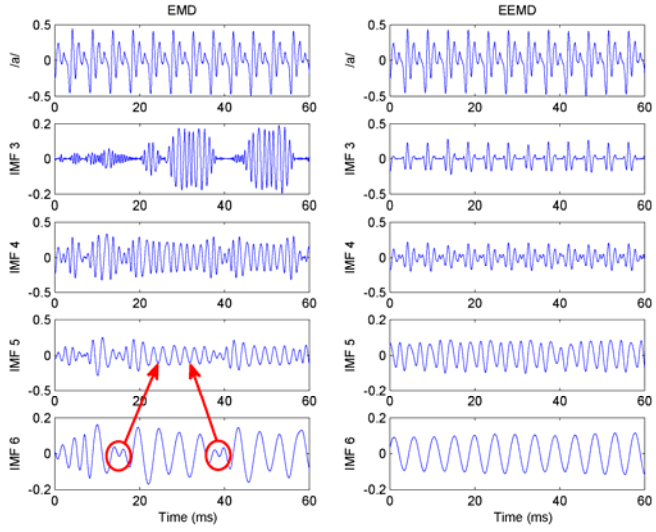

Fig. 1 Sustained vowel /a/ analyzed by EMD (left column) and EEMD (right column). IMFs 4 to 6 are shown. In the left column, the circles indicate two segments where "mode mixing" occurs.

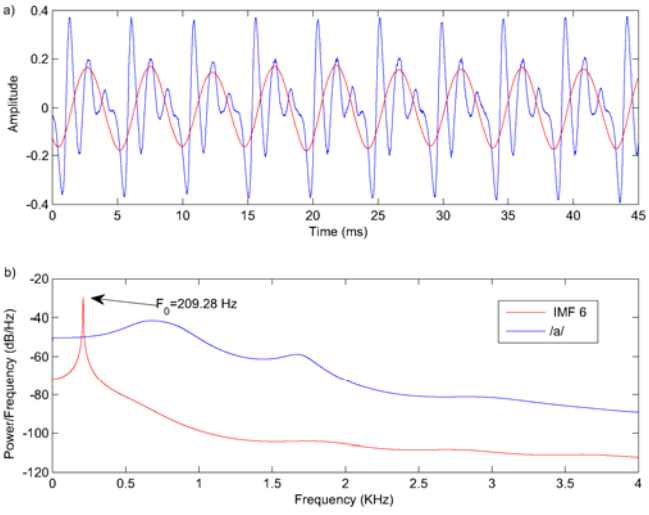

Fig. 2 a) Sustained vowel /a/ (blue) and IMF6, obtained by EEMD (red). b) PSD of the sustained vowel /a/ (blue) and its EEMD based IMF6 (red). The peak of the spectrum of the IMF6 is marked as F_0 = 209.28 Hz.

and (ii) at any point, the mean value of the upper and lower envelopes is almost zero. Given a signal *x(t)*, the non-linear EMD algorithm is described as follows [3]:

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 1.
   - if *d(t)* is an IMF, evaluate the residue $r(t) = x(t) - d(t)$.
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]. One of the most significant EMD drawbacks is the so called mode mixing, illustrated in the left column of Fig. 1, where a sustained vowel /a/ is analyzed by EMD. Only the four IMFs with higher energy are shown. It is clear the appearance of oscillations of quite different scales in IMF3. Another example can be observed in IMF6, where two oscillations, very similar to those on IMF5, are marked with circles.

The EEMD algorithm [7] alleviates the mode mixing defining the true IMF components as the mean of certain ensemble of trials, each one obtained by adding Gaussian white noise of finite variance to the original signal.

In the right column of Fig. 1 an example of the EEMD abilities can be seen, obtained with an ensemble of size *Ne* = 5000 and different additive noise with standard deviation ε = 0.2. The IMFs 3 to 6 are shown. Compared with the left column, they appear to be more regular. The fundamental period of the sustained vowel /a/ is captured by IMF6 as can be appreciated in Fig. 2.a, where the analyzed signal and IMF6 are shown. In Fig. 2.b the power spectral densities (PSD) of /a/ and IMF6 are plotted. The PSD of IMF6 have a well defined peak in the frequency F = 209.28 Hz. It can be understood as the mean fundamental frequency.

C. EEMD based F_0 extraction algorithm

In this section the main ideas for the EEMD based F_0 extraction algorithm are presented and discussed.

Once the EEMD is computed, a first question arises about which mode holds F_0. With this in mind, we perform a visual inspection to the decomposition of the normal voices in our database and, as above, we select a candidate mode. With the purpose of eliminating spurious frequency components, a band-pass Chebyshev Type II filter is applied to the selected mode. The filter is centered in the frequency corresponding to the maximum of the PSD of the selected mode. As shown in Fig. 2.b, this frequency is a good approximation to the mean of the F_0. We use a 150 Hz filter bandwidth. Then, an AM – FM separation algorithm must be applied. We select DESA-1[10], which overcomes the Hilbert transform based techniques when applied to real-world signals [11].

In the case of our 53 normal sustained vowels, F_0 was found in the fifth, sixth or seventh IMF. In two occasions F_0 was found in IMF7, with averages 120.394 Hz and 121.102 Hz; nineteen times in IMF6, with averages between 121.652 Hz and 189.295 Hz; and in IMF5 in the 32 remaining voic-

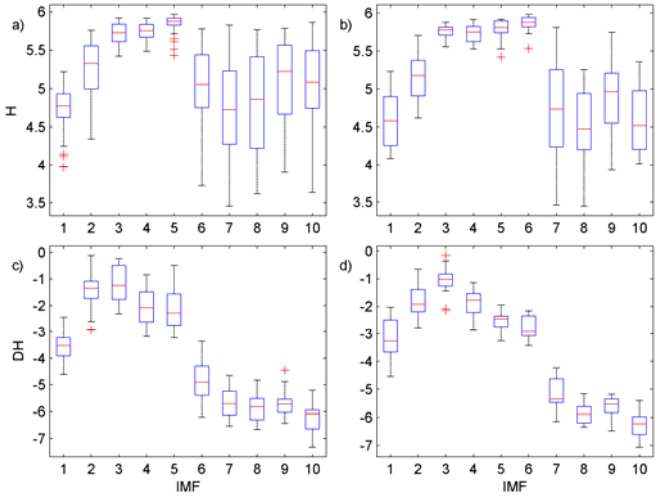

Fig. 3 a, b) Boxplots of the discrete entropy of IMFs 1 to 10 of normal sustained vowels /a/ in which F_0 is present in IMF5 and IMF6, respectively. c, d) Boxplots of the differential entropy of IMFs 1 to 10 of normal sustained vowels /a/ in which F_0 is present in IMF5 and IMF6, respectively

es, averaging between 193.934 Hz and 316.504 Hz. It can be appreciated that the IMFs containing the F_0 depends on its mean value. It is also in agreement with the studies of Flandrin et al., who showed that the EMD is an adaptive dyadic filter bank when applied to white noise [12].

An automatic method for selecting the mode where the F_0 is present is needed. We explore the ability of entropies in this task, as information related measures. In Fig. 3.a, the box-plots of the discrete Shannon entropies (H) [13], estimated with 500 bins, for the ten first modes of the sustained vowels /a/ in which F_0 was encountered in IMF5, are shown. In Fig 3.b those corresponding to the voices in which F_0 was encountered in IMF6 are shown. The first mode mainly contains the noise residuals (the one added and the one present in the original signal). It has a lower entropy than the next four (Fig 3.a) or five (Fig 3.b) modes. In case of voices in which F_0 is in IMF5, the entropy has a step in mode 5, while there is a similar step in mode 6 for voices in which F_0 is in IMF6. These steps have a lower overlap while using the differential entropy (DH) [13], also with 500 bins. The results are shown in Fig. 3.c (F_0 in IMF5) and Fig. 3.d (F_0 in IMF6). Here it must be emphasized that the IMFs obtained by EEMD for normal sustained vowels, have a sinusoid-like morphology. Indeed, their probability density functions are also similar. The DH of a sinusoid with amplitude A is given by $\mathrm{DH} = \ln(\pi A/2)$. Therefore, it is reasonable to think that the logarithm of the IMFs' power could be also a good index for finding the mode which holds F_0. This approach will be addressed in other works.

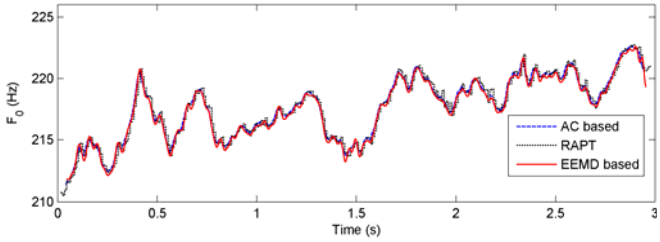

Fig. 4 F_0 of a healthy sustained vowel /a/. Autocorrelation method (black), RAPT (blue) and the EEMD based instantaneous F_0 (red).

Based on these results, for each mode = 5, 6, and 7, for normal voices we can propose the thresholds T_5, T_6, and T_7 as the followings: $-3.365 < T_5 < -3.234$, $-4.224 < T_6 < -3.433$, and $-5.761 < T_7 < -4.172$. In this way, if the DH of IMF5 is higher than T_5 while the one of IMF6 is lower than T_5, then F_0 is expected to be in IMF5. Otherwise, we test the existence of a step between IMF6 and 7 using T_6, and next between IMFs7 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.

III. Results

For illustration purposes, the F_0 extracted with the method proposed in the previous section from a sustained vowels /a/ in the database is presented in red in Fig. 4. For comparison, two additional pitch extraction methods – RAPT (black) [14], and an autocorrelation-based method (blue) [15] – are applied to the same normal voice records and also shown in Fig. 4. The parameters involved in these two algorithms are the defaults. It can be observed that the results are similar, although it must be remembered the stair-case nature of the two last methods. The Pearson correlation coefficient between the mean F_0 of the 53 healthy sustained vowels /a/ reported in [8] and the averaged instantaneous frequency obtained by our method was r = 0.999995.

In Fig. 5 the F_0 of a sustained vowel /a/ from a patient suffering muscular tension dysphonia, obtained with the proposed method, is shown. As in Fig. 4, the F_0 obtained with RAPT and auto-correlation based methods are superposed in blue and black. Even if the autocorrelation based method has been reported as being the best pitch estimation technique for the analysis of pathological sustained vowel /a/ [16], it can be observed in Fig. 5 that it fails. Also does RAPT algorithm, while the method here proposed exhibits a 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,

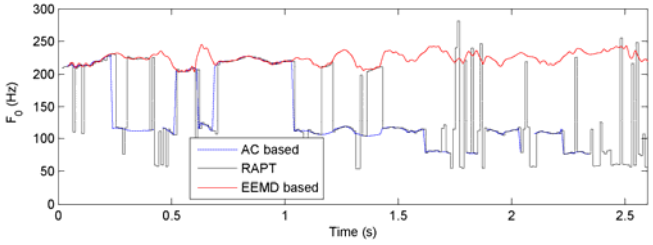

Fig. 3 F_0 of a pathological sustained vowel /a/. Autocorrelation method (black), RAPT (blue) and EEMD based instantaneous F_0 (red).

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 appear. In the method here proposed, we have observed that these spike-like artifacts were coincident with pathological voice segments 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 evident. 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 propose 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.

IV. Discussion and Conclusions

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 algorithm (DESA-1). Additionally, an entropy based 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. These preliminary results suggest that the method here proposed provides important improvements to this task 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. 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. M. C. Jackson–Menaldi, of Lakeshore Professional Voice Center of the Lakeshore Ear, Nose and Throat Center, St. Clair Shores (USA) and Wayne State University, Detroit (USA), for her valuable suggestions. The authors also thank Kay Elemetrics Corp.

References

1. Hess W (2008) Pitch and voicing determination of speech with an extension toward music signals. In Benesty M, Sondhi M, Huang Y (Eds.) Springer handbook of speech processing. Springer-Verlag
2. Schlotthauer G, Torres ME, Jackson-Menaldi MC (2009) A pattern recognition approach to spasmodic dysphonia and muscle tension dysphonia automatic classification. J Voice (In press)
3. Huang N et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc B 454:903–995
4. Dimitriadis D, Maragos P (2006) Continuous energy demodulation methods and application to speech analysis. Speech Comm 48:819–837
5. Huang H, Pan J (2006) Speech pitch determination based on Hilbert-Huang transform. Signal Process 86:792–803
6. Weiping H, Xiuxin W, Yaling L, Minghui D (2005) A novel pitch period detection algorithm bases on HHT with application to normal and pathological voice, Proc. IEEE Eng. Med. and Biol. 27th Annual Conf, Shangai, China, 2005, pp 4541–4544
7. Wu Z, Huang N (2009) Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv Adapt Data Anal 1:1–41
8. Kay Elemetrics Corp (1994) Disordered voice database 1.03, Massachusetts Eye and Ear Infirmary, Voice and Speech Lab, Boston, USA
9. Rilling G, Flandrin P, Gonçalvès P (2003) On empirical mode decomposition and its algorithms, Proc. IEEE–EURASIP Workshop NSIP-03, Grado, Italy, 2003
10. Maragos P, Kaiser J, Quatieri T (1993) Energy separation in signal modulations with application to speech analysis. IEEE Trans Signal Process 41:3024–3051
11. Díaz M, Esteller R (2007) Comparison of the non linear energy operator and the Hilbert transform in the estimation of the instantaneous amplitude and frequency. Lat Am Trans IEEE 5:1–8
12. Flandrin P, Rillings G, Gonçalvès P (2004) Empirical mode decomposition as a filter bank. IEEE Signal Process Lett 11:112–114
13. Papoulis A (1991) Probability, random variables and stochastic processes, third edition, McGraw-Hill
14. Talkin D (1995) A robust algorithm for pitch tracking (RAPT). In Kleijn W, Paliwal K (Eds.) Speech Coding & Synthesis. Elsevier
15. Boersma P (1993) Accurate short–term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc. Institute of Phonetics Sciences 17:97–110
16. Jang S et al (2007) Evaluation of performance of several established pitch detection algorithms in pathological voices, Proc. 29th Annual Ing. Conf. IEEE-EMBS, Lyon, France, 2007, pp 620–623