Empirical mode decomposition. Spectral properties in normal and pathological voices

M.E. Torres[1,2], G. Schlotthauer[1], H.L. Rufiner[2,3] and M.C. Jackson-Menaldi[4]

[1] Lab. de Señales y Dinámicas no Lineales, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde, Argentina.
[2] Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Oro Verde, Argentina.
[3] Lab. de Señales e Inteligencia Computacional, Fac. de Ing. y Cs. Hídricas, Universidad Nacional del Litoral, Sta. Fe, Argentina.
[4] Dept. of Otolaryngology, School of Medicine, Wayne State University, Detroit, Michigan, USA.

Abstract— **Empirical Mode Decomposition is a data driven technique proposed by Huang. In this work, we explore spectral properties of the intrinsic mode functions and apply them to speech signals corresponding to real and simulated sustained vowels. For the synthetic sustained vowels we propose a phonation model that includes perturbations implied in common laryngeal pathologies. We extract features from each signal using the Burg's standard spectral analysis of their intrinsic mode functions. Due to its well-known theoretical properties, the classic K-nearest neighbor's classification rule is applied to real and synthetic data. We show that even using this basic pattern classification algorithm, the selected spectral features of only three intrinsic mode functions are enough to discriminate between normal and pathological voices. We have obtained a 99.00% of correct classifications between normal and pathological synthetic voices (K=1, sensitivity=0.990, specificity=0.990); while in the case of real voices the percentage of correct classification was 93.40% (K=3, sensitivity=0.925, specificity=0.926). These results strongly suggest that spectral properties of Empirical Mode Decomposition provide useful discriminative information for this task. Additionally we consider two pathologies of different etiology and treatment, which, given the similarity of their voice characteristics, are frequently misdiagnosed in clinical practice: muscular tension dysphonia and adductor spasmodic dysphonia. Preliminary results with a reduced real data base suggest that this approach could provide useful orientation to physicians and voice pathologists.**

Keywords— **Empirical mode decomposition, speech analysis, pathological voices, spectral analysis.**

I. INTRODUCTION

Empirical Mode Decomposition (EMD) has been recently proposed by Huang *et al.* [1] for adaptively decomposing 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 biological applications [1, 2, 3] and interpretations [4, 5]. The method consists in a local and fully data-driven splitting of a (possibly non-stationary) signal in fast and slow oscillations.

In this work, we explore some spectral properties of the IMFs. The comparison with the spectra of real data and its IMFs allows us to present preliminary results of an application of this method to the analysis and discrimination between normal and pathology-cal speech signals. We also study a couple of dysphonias: Adductor Spasmodic Dysphonia (AdSD) and Muscular Tension Dysphonia (MTD), two voice disorders with different etiology [6], frequently confused and not easily identified by local clinicians.

The primary treatment for MTD is voice therapy, which is only of limited benefit to patients with AdSD when used as a sole treatment modality. Although these disorders have been described in the literature, the symptoms have not been well defined and may appear similar, and those of AdSD might be confused with those of essential vocal tremor or of muscle tension dysphonia (MTD).

A recent review on symptoms for AdSD and MTD [7] confirm multifactorial etiologies contributing to hoarseness in the patients identified with MTD, concluding that an interdisciplinary approach to treating all contributing factors portends the best prognosis. Therefore, patients might not be easily identified by local clinicians for treatment [8]. In spite that a classical Fourier-based analysis would be useful to detect hoarseness, it should not provide good information given that this symptom can be present in both pathologies. Therefore, it is mandatory to provide appropriate new tools that could help physicians and voice pathologists to discriminate between these two pathologies.

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 [9], 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.

J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen (Eds.): ECIFMBE 2008, IFMBE Proceedings 22, pp. 252–255, 2008
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2009

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. II basic concepts to be used are described. In Sec. III materials are described. In Sect. IV we present the results. Finally, in Sec. V conclusions are presented.

II. Basic concepts

The Empirical Mode Decomposition (EMD) is a method developed to deal with data from nonstationary and nonlinear processes [1]. The decomposition is based on the intuitive assumption that any data consists of different simple intrinsic modes of oscillations. Each intrinsic mode, linear or nonlinear, represents a simple oscillation, which will have the same number of extrema and zero-crossings. Furthermore, the oscillation will also be symmetric with respect to the "local mean". Each of these oscillatory modes is represented by an IMF with the following definition [2]: (i) in the whole data set, the number of extrema and the number of zero-crossings must be either equal or differ at most by one, and (ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. Therefore a signal *x(t)* can be decomposed as follows [1]:

1. Identify all its local extrema.
2. Interpolate between minima (resp. maxima), obtaining an "envelope" $e_{\min}(t)$ (resp. $e_{max}(t)$.)
3. Compute the local trend $r(t)=(e_{\min}(t)+ e_{\max}(t))/2$
4. Extract the local detail $d(t)=x(t)-r(t)$
5. Iterate 1-4 on r(t), until some stopping criterion.

The above procedure has to be refined by a *sifting* process [4]. The obtained local details d(t) are the IMFs.

III. Materials

We have used real and simulated voices for our experiments. Signals have been Hamming windowed to perform the EMD analysis and the formants estimation.

Simulated voices. In order to explore the discrimination power of the proposed technique, experiments with normal and pathological synthetic voices have been carried out. These signals have been generated using a phonation model that incorporates the perturbations implied 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*:

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

where P_0 is the mean period and $\sigma_P = P_0\, jitt\% \big/ 200$. 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:

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

Taking into account statistics obtained with real signals, we have simulated 400 signals, corresponding to 100 male and 100 female, for each group of normal and pathological voices. We adopted a fundamental frequency with a distribution *N(144,22.5)* for male voices and *N(245,24.5)* for female voices; a *N(0.4,0.1)* jitter distribution for normal voices and *N(5,1)* for pathological voices; and a shimmer with distribution *N(1,0.2)* and *N(8,1)* respectively.

Real voices. 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 has been used. The data was then low-pass filtered and down-sampled to 22 KHz. All the voices were classified by an experienced voice pathologist.

We considered a first set of 106 voice (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

J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen (Eds.): ECIFMBE 2008, IFMBE Proceedings 22, pp. 252–255, 2008
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2009

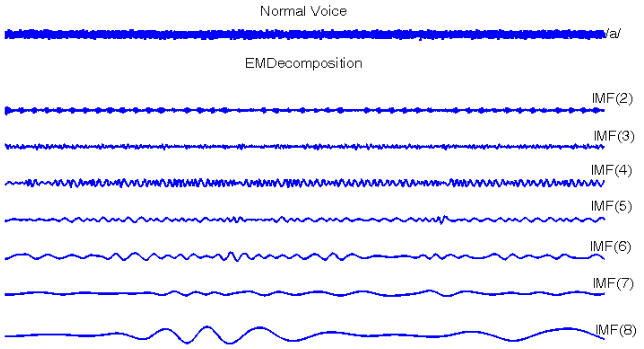

Fig. 1 Sustained vowel /*a*/ in the first row and IMFs 2-8 of its EMD.

Data Base DB2. Patients affected with AdSD may attempt to prevent their symptoms by increasing the tension in their laryngeal muscles in an effort to compensate them. The consequence is the appearance of additional physical disturbances similar to MTD along with AdSD. The overriding symptoms of MTD can escalate over time making difficult to discern the underlying symptoms of AdSD [8].

IV. Results

The EMD algorithm of the vowels stopped in average at IMF 12±1. As an example a sustained normal vowel /*a*/ and the first eight IMFs of its EMD are shown in Fig.1. Inspired by Fig. 2 here we propose to consider the maxi-mum Psd of the IMFs 2-4 and the corresponding frequencies as new features to be used for our classification purposes.

A. Simulated normal and pathological voices

In order to study the classification capability of the new tools here presented, we have constructed the feature vectors with the maximum Psd (log2) and the corresponding frequencies, of IMF(i), i=2,3,4, for each of the simulated voices. The 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.

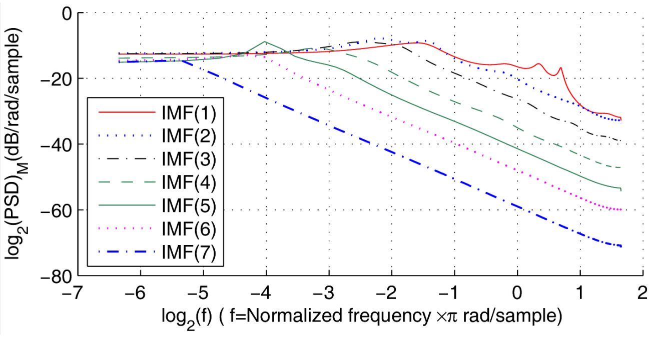

Fig. 2 Log-log power spectrum density, estimated with Burg algorithm, corresponding to each of the IMFs of a Spanish sustained vowel /*a*/

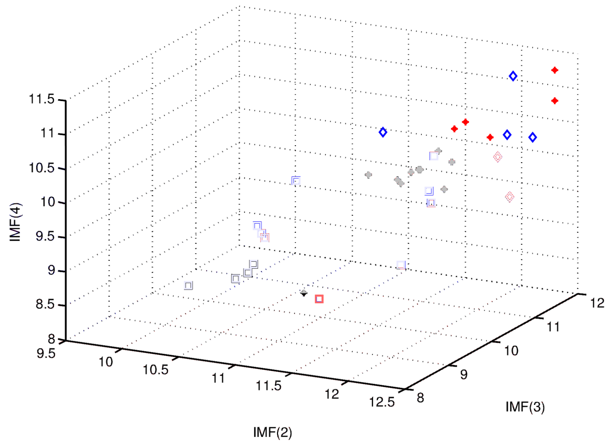

Fig. 3: Frequency (log2) corresponding to the maximum Psd of three IMFs, of normal (stars) and pathological (diamonds – MTD – and squares – ASD) voices (DB2).

With this simple and general-purpose classifier, the best performance has been obtained using K=1, reaching a 99% of correct classifications. In Table 1.a we present the obtained confusion matrix. This result confirms that the IMFs' spectrum provides 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 to short-term perturbations of their fundamental frequency and intensity. Therefore, the proposed method is able to distinguish between normal and altered voices with very similar Fourier spectra.

B. Real normal and pathological voices

Following the same procedure as in Sec. IV, with the real voices DB1 described in Sec. III., 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 have 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 results 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. 3 that it seems to be possible to separate AdSD from the normal and MT. Plotting the maxima of the Psd (in log2), we see in Fig. 4 that it is possible to separate most of the MTD from the other pathology and the normal ones.

V. Conclusions

In this paper we have 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%). In the case of discrimination between normal and pathological real voices we have obtained a performance of (93.40%). We consider that it could be possible to overcome the best reported value by refining the proposed method.

The 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.

Table 1 : Confusion matrix

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

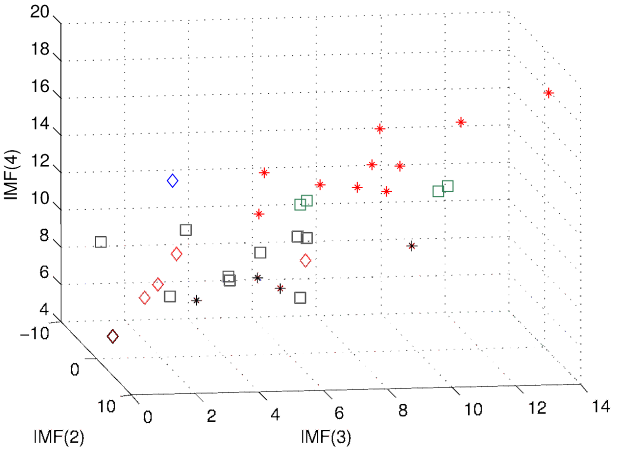

Fig. 4 : Maxima Psd (log2) corresponding to three IMFs of normal (stars) and pathological (diamonds – MTD – and squares – ASD) voices (DB2)

Acknowledgment

This work was supported by ANPCyT under project PICT 11-127-00, UNER and CONICET.

References

1. Huang NE, Shen Z, Long SR et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond A. 454:903–995
2. Huang NE, Shen SSP Eds (2005) Hilbert-Huang transform and its applications, Vol. 5. World Scientific
3. Schlotthauer G., Torres ME (2005) Descomposición modal empírica: análisis y disminución de ruido en señales biológicas, Proc. XV Congreso Argentino de Bioingeniería SABI'2005, Paraná, E.R. Argentina, 2005, ISBN 950-698-155-8. File:101PS.pdf
4. Rilling G, Flandrin P, Gonçalvès P (2003) On empirical mode decomposition and its algorithms, Proc IEEE-EURASIP Workshop on Nonlinear Signal and Image Process., NSIP-03, Grado (I)
5. Rilling G, Flandrin P (2006) On the influence of sampling on the empirical mode decomposition Proc. Acoustics, Speech and Signal ICASSP 2006, DOI 10.1109/ICASSP.2006.1660686
6. Verdolini K, Rosen CA, Branski RC Eds.(2006) Special Interest Division 3. Voice and Voice Disorders. American Speech-Language-Hearing Association, Classification Manual for Voice Disorders - I. Lawrence Erlbaum Associates, Inc
7. Atkinson C., Altman KW, Lazarus C (2005) Current and emerging concepts in muscle tension dysphonia: A 30-month review. J. Voice, 19: 261–267
8. Jackson-Menaldi MC (2002) La Voz Patológica, Editorial Médica Panamericana, Buenos Aires
9. Schlotthauer G, Torres ME, Jackson-Menaldi MC (2006) Automatic Classification of Dysphonic Voices. WSEAS Trans. Signal Process., 2:1260-1267. (And references therein)

J. Vander Sloten, P. Verdonck, M. Nyssen, J. Haueisen (Eds.): ECIFMBE 2008, IFMBE Proceedings 22, pp. 252–255, 2008
www.springerlink.com © Springer-Verlag Berlin Heidelberg 2009