ESTIMATION OF SPECTRAL ENVELOPE AND HARMONIC COMPONENT IN SPEECH USING EMPIRICAL MODE DECOMPOSITION

M. Kemiha[1] and A. Kacha[2]

[1]*Département d'Electronique, University of Jijel, Algeria*
[2]Laboratoire de Physique de Rayonnement et Applications, University of Jijel, Algeria
Email : kemihamina@yahoo.fr, akacha@ulb.ac.be

ABSTRACT

Empirical mode decomposition (EMD) algorithm is proposed as an alternative to decompose the magnitude spectrum of the speech signal into harmonic and spectral envelope components. The EMD is a tool for the analysis of multi-components signals. The analysis method does not require a priori fixed basis function like conventional analysis methods (e.g. Fourier transform and wavelet transform). The EMD algorithm decomposes adaptively a given signal into oscillation modes namely the intrinsic mode functions (IMFs) extracted from the signal itself. An adaptive method is developed to select the IMF index that enables to separate the harmonic component and the spectral envelope and then the IMF variance is used as a parameter to perform clustering in order to distinguish the IMFs that constitute the harmonic component and those constituting the spectral envelope. The proposed method is tested on both synthetic and natural speech signals.

Index Terms— empirical mode decomposition, intrinisic mode function, speech decomposition.

1. INTRODUCTION

In speech production, speech sounds are considered as produced by the action of a filter which models the vocal tract on an acoustic source [1][2]. This model is called source-filter model. The acoustic source is a periodic sequence for voiced sounds and a random noise sequence for unvoiced sounds. The magnitude spectrum of a voiced speech signal can be viewed as the combination of two components: a slowly varying component that results from the contribution of the vocal tract and a rapidly varying periodic component which is the effect of the periodic source (excitation).
The decomposition of the magnitude spectrum of the speech signal into its vocal tract and source contributions plays a central role in many areas of speech processing such as pitch and formants estimation, speech synthesis, speech enhancement, speech and speaker recognition, etc. Given a speech spectrum, the aim is to separate the effect of the vocal tract from its excitation at the glottis. The separation of these two components is very attractive for voice processing because it provides a way to analyze, study and understand the properties of voice production. In speech and speaker recognition, the deconvolution is carried out in order to estimate the formant frequencies which are the effect of the eigenmodes that characterize the vocal tract [1]. In medical applications of speech processing, the filtering effect of the vocal tract has to be eliminated and the parameters of the excitation are used to derive objective measures for quality assessment of voice of dysphonic speakers [2][3]. In addition, a voice can be transformed and synthesized using independent manipulation of its elements.
Despite the large number of methods proposed in the literature, the estimation of the spectral envelope and the harmonic component is most often treated as two independent problems. If the harmonic component (fundamental frequency) is known, the spectral envelope can be estimated reliably and conversely if the spectral envelope is known, the harmonic component can be estimated accurately. Conventional methods for estimating spectral envelope are based on linear predictive coding (LPC) [1] or real cepstrum [3]. The estimation of the harmonic component or the fundamental frequency (pitch) is based on parameters that exploit the assumption of local periodicity of voiced speech sounds either in the time domain or in the spectral domain [5][6].
A method that has been proposed as an alternative to estimate simultaneously the spectral envelope and the harmonic component is the wavelet-based deconvolution approach [7]. Experiments have shown that the wavelet-based method for speech separation provides satisfactory results with the data frame length more than or equal to 1024 samples. A drawback of the wavelet decomposition method for speech deconvolution is the use of an a priori given basis functions making this approach non-optimal for all kinds of speech signals.
Recently, a signal decomposition method, called empirical mode decomposition (EMD), has been introduced for analyzing data from nonstationary and/or nonlinear processes [8]. The EMD has received more attention in terms of applications, interpretation and improvement. The major advantage of the EMD is that the basis functions are derived from the signal itself and not fixed a priori.
In this paper, the EMD is proposed as an alternative to separate the harmonic component and the spectral envelope of the speech signal. The proposed method of separation operates in the log-spectral domain. The effectiveness of the proposed approach is evaluated on both synthetic and real speech and its performance is compared to that of the wavelet-based separation method. The remainder of the

paper is organized as follows. Empirical mode decomposition algorithm is introduced in Section 2. The EMD-based approach for speech components separation is presented in Section 3. Results based on both synthetic and real speech signals are presented in Section 4. Finally, conclusions are given in Section 5.

2. EMPIRICAL MODE DECOMPOSITION

The empirical mode decomposition has been proposed by Huang et al. as a new signal decomposition method for nonlinear and/or nonstationary signals [8]. The EMD decomposes a given signal into a collection of oscillatory modes, called intrinsic mode functions (IMFs), which represent fast to slow oscillations in the signal. Each IMF can be viewed as a sub-band of the signal. Therefore, the EMD can be viewed as sub-band signal decomposition. Conventional signal analysis tools, such as Fourier or wavelet-based methods, require some predefined basis functions to represent a signal. The EMD relies on a fully data-driven mechanism that does not require any a priori known basis. The algorithm operates through the following steps :

1. Initialize the algorithm: j=1, initialize residue $r_0(t)$=$x(t)$ and fix the threshold δ
2. Extract local maxima and minima of r_{j-1}(t)
3. Compute the upper envelope $Uj(t)$ and lower envelope $L_j(t)$ by cubic spline interpolation of local maxima and minima, respectively
4. Compute the mean envelope $m_j(t) = \left(U_j(t) + L_j(t)\right)/2$
5. Compute the jth component $h_j(t)$=r_{j-1}(t) - $m_j(t)$
6. $h_j(t)$ is processed as $r_{j-1}(t)$. Let $h_{j,0}(t)$=$h_j(t)$ and $m_{j,k}(t)$, k=0, 1, …, be the mean envelope of $h_{j,k}(t)$, then compute $h_{j,k}(t)$=$h_{j,k-1}(t)$-$m_{j,k-1}(t)$ until

$$SD_k = \sum_{t=0}^{T} \frac{\left|h_{j,k-1}(t) - h_{j,k}(t)\right|^2}{\left(h_{j,k-1}(t)\right)^2} < \delta$$

7. Compute the jth IMF as $IMF_j(t)$=$h_{j,k}(t)$
8. Update the residue $r_j(t)$=$r_{j-1}(t)$ - $IMF_j(t)$
9. Increase the sifting index and repeat steps 2 to 8 until the number of local extrema in $r_j(t)$ is less than 3

The signal reconstruction process is given by (1), which involves the IMFs and the residual obtained via the EMD algorithm:

$$x(t) = \sum_{j=1}^{N} IMF_j(t) + r_N(t) \quad (1)$$

3. EMD-BASED SPEECH COMPONENTS SEPARATION

According to the source-filter model of speech production, voiced speech is the effect of the convolution of the excitation of the vocal tract system and its impulse response, so that we may assume the following relationship [1]

$$x(t) = e(t) * v(t) \quad (2)$$

where $x(t)$ is the speech signal, $v(t)$ is the impulse response of the vocal tract system, and $e(t)$ is the excitation signal which originates at the vocal cords, and * denotes the convolution. Windowing the signal frame $x(t)$ and taking the Fourier transform magnitude gives

$$\left|X_w(f)\right| = \left|E_w(f) \times V(f)\right| \quad (3)$$

where f denotes the frequency, $X_w(f)$, $E_w(f)$ are short-time magnitude spectrum of the windowed speech frame and windowed excitation signal, respectively and $V(f)$ is the frequency response of the vocal tract.

Taking the logarithm changes the multiplicative components into additive components.

$$\log\left|X_w(f)\right| = \log\left|E_w(f)\right| + \log\left|V(f)\right| \quad (4)$$

It is observed that the log magnitude spectrum is the sum of two spectral components, $\log\left|E_w(f)\right|$, the log magnitude spectrum of the windowed excitation signal and $\log\left|V(f)\right|$, the spectral envelope. The log magnitude spectrum can be considered as composed of a slowly varying (with respect to frequency) contour due the contribution of the vocal tract and a series of harmonics characterized by a periodic structure. The EMD algorithm yields an effective tool that enables to separate the two components of the log magnitude spectrum indeed, the EMD algorithm acts as a filterbank [9], so that the decomposition of the log magnitude spectrum via the EMD algorithm results into several oscillating components (IMFs) that can be clustered in two classes where each class of components is associated to some part of the log magnitude spectrum.

It has been shown that the IMF variance for speech signals significantly decreases after the fourth IMF, as the IMF order increases [10]. It was found experimentally that the IMF statistics for a speech signal are characterized by a peak IMF energy in a higher IMF order.

This IMF variance build-up is used to select the IMF order, to use in the speech components reconstruction. Therefore, The IMF index is determined by examining the trough in prior to each identified peak. The method used to select the optimal index that enables to separate the harmonic component and the spectral envelope follows that described in [10] :

1) Compute the variance $V(m)$ of the mth IMF as

$$V(m) = \frac{1}{L}\sum_{f} IMF_m^2(f) \quad (5)$$

L is the length of the IMFs

2) Identify the indices of the peaks, $\mathbf{m}_p$ in $V(m)$ for $m > 4$

3) Find the indices of the troughs $\mathbf{m}_t$

4) Compute the IMF variance build-up $\mathbf{m}_b$ to those peaks using $\mathbf{m}_b$=$\mathbf{m}_p$ - $\mathbf{m}_t$

5) Determine the index i of the first occurrence of the largest build-up $m_{b,i}$ in $\mathbf{m}_b$ and select the corresponding peak $m_{p,i}$ in $\mathbf{m}_p$

6) Determine the IMF index M as $M=m_{p,i} - m_{b,i}$
The different IMF_j, j=1, 2, …, N are clustered as follows :
If $j<M$, IMF_j belongs to the harmonic component
If $j\geq M$, IMF_j belongs the spectral envelope

The harmonic component and the spectral envelope are estimated, respectively, as

$$\log|E_w(f)| = \sum_{j=1}^{M-1} IMF_j(f) \quad (6)$$

$$\log|V(f)| = \sum_{j=M}^{N} IMF_j(f) + r_N(f) \quad (7)$$

To better understand the proposed method, the log magnitude spectrum of the signal windowed by a 1024-samples Hamming window is illustrated in Fig. 1. The empirical mode decomposition of the log magnitude spectrum results in N=4 IMFs as shown in Fig.1 (a). The IMF variance versus the IMF order is displayed in Fig. 1 (b). As it is observed, the peak is achieved at m_p=4 and the trough is m_t=2 giving a variance build-up $m_b=m_p - m_t$=2 and IMF index M=2. According to the clustering algorithm, the harmonic component is estimated as the first IMF , and the spectral envelope is obtained as the sum of the remaining three IMFs and the residue.

4. RESULTS AND DISCUSSION

The proposed approach has been tested on synthetic speech signals as well as on natural speech and its performance in terms of accuracy has been compared to that of the wavelet-based separation method. The sampling rate of all speech signals used in the experiment is 20 kHz. The artificial signal used in the test is a 1-second synthetic vowel /a/ generated according to the source-filter model of speech production.
The source-filter model consists of a source that generates a periodic impulse train to model glottal airflow and a vocal tract modeled as an all-pole filter characterized by three poles [1][11] corresponding to the formant frequencies 981.6 Hz, 1631.3 Hz and 3159.9 Hz and bandwidths 140 Hz, 180 Hz and 55 Hz, respectively. Lip radiation is modeled by a first order difference operator $R(z)=1-z^{-1}$.

Conversely, the spectral envelope estimated via the wavelet-based separation approach has lost too much details compared to the true model for both frame lengths.

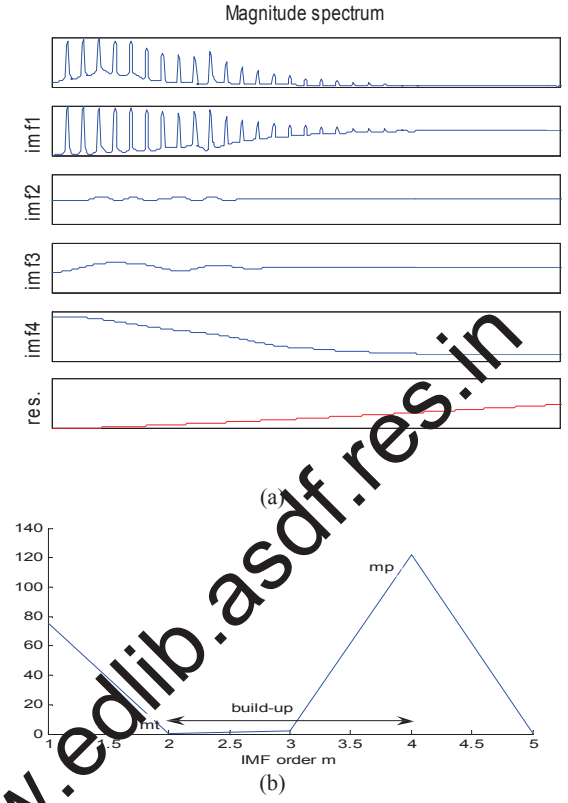
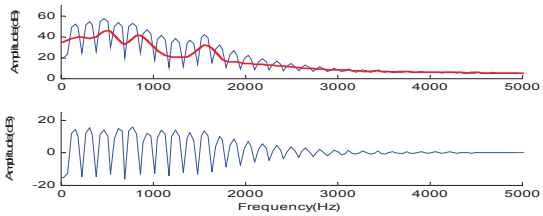
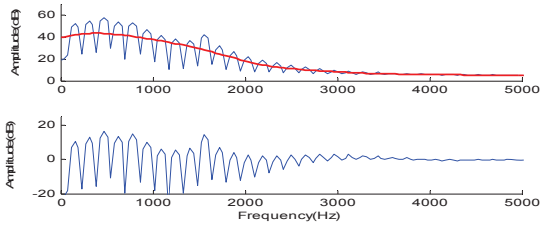


Fig. 1: Illustration of the separation of the harmonic component and spectral envelope of synthetic /a/ via empirical mode decomposition. (a) Log magnitude spectrum and IMF components. (b) IMF variances.

The speech signal has been divided into K non-overlapping frames and the harmonic component and the spectral envelope have been computed for each frame using the wavelet-based method and the EMD-based technique for different frame sizes. Fig. 2 displays the average harmonic component and the average spectral envelope estimated by using frame lengths of 256 and 1024 samples. As can be seen, the wavelet-based method fails to estimate accurately the harmonic component and the spectral envelope from the log magnitude spectrum. The EMD-based approach provides accurate estimates for both components whatever the frame length. Fig. 3 shows the spectral envelope estimated via both methods for frame lengths of 256 and 1024. The estimates have been superimposed to the true model. As can be observed, the EMD-based approach results in formant frequencies located with high accuracy.

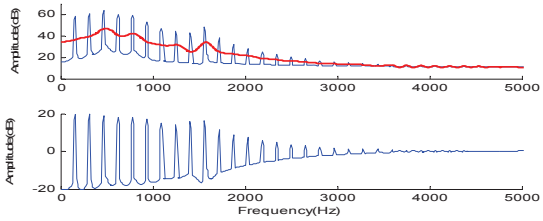


(i) EMD-based separation

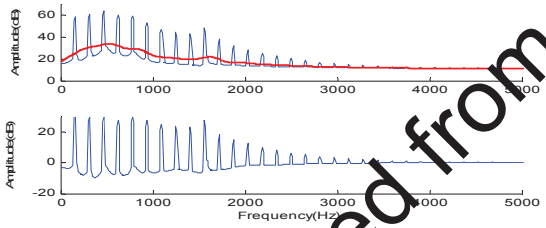


(ii) Wavelet-based separation

(a)



(i) EMD-based separation

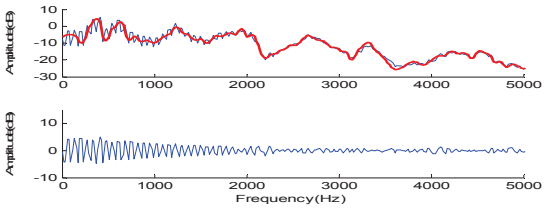


(ii) Wavelet-based separation

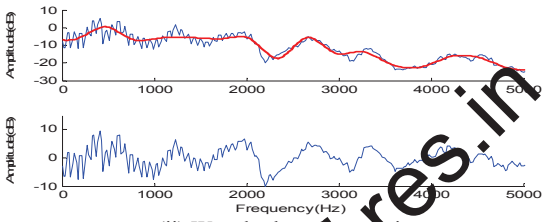
(b)

Fig. 2: Comparison between wavelet-based and EMD-based separation methods applied to a synthetic vowel /a/ for different frame lengths. (a) Frame length of 256. (b) Frame length of 1024.

The EMD-based method and the wavelet-based technique have been applied to natural vowel /a/ produced by a male speaker. Fig. 3 displays the harmonic component and the spectral envelope estimated via both methods by using frame lengths of 256 and 1024 samples. As illustrated, for a frame length of 256, the wavelet-based separation method results in a smooth estimate of the spectral envelope and in a harmonic component that contains slow variations due to the contribution of the vocal tract. For a frame length of 1024, the harmonic component is accurately estimated but the formant frequencies are still undistinguishable from the spectral envelope. The EMD-based separation method yields accurate estimates of the harmonic component and spectral envelope for both frame lengths.

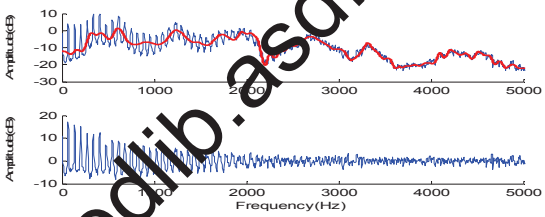


(i) EMD-based separation

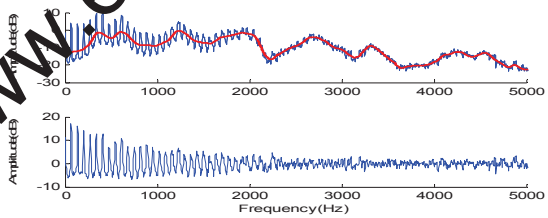


(ii) Wavelet-based separation

(a)



(i) EMD-based separation

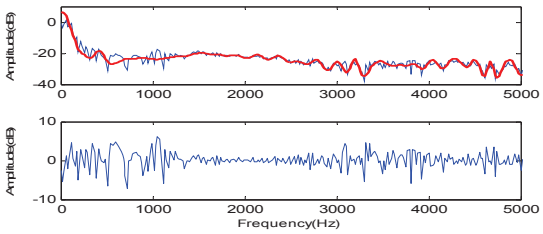


(ii) Wavelet-based separation

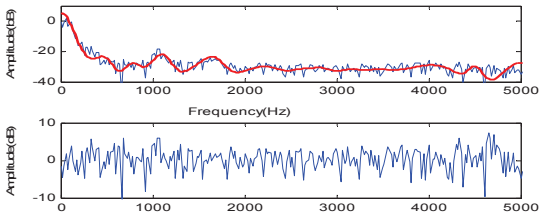
(b)

Fig. 4: Comparison between wavelet-based and EMD-based separation methods applied to a natural vowel /a/ for different frame lengths. (a) Frame length of 256. (b) Frame length of 1024.

The proposed method has been applied to the separation of the spectral envelope and harmonic component in continuous speech. As an illustration, Fig. 5 shows the separation results obtained via the EMD-based approach and the wavelet-based method for two successive voiced frames extracted from the speech signal corresponding to a sentence uttered by a male speaker. The frame length has been fixed to 256. As can be observed, the EMD-based method provides more accurate estimate of the harmonic component than that obtained with the wavelet-based approach. The wavelet-based approach is unable to reveal the spectral peaks corresponding to the formants whereas the EMD-based approach distinguishes clearly the different formants.

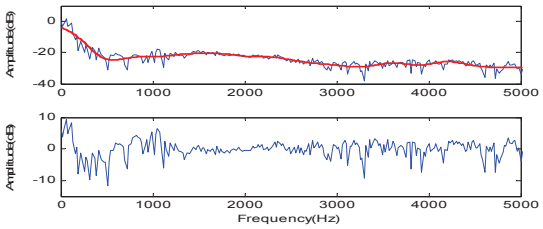


(i) EMD-based separation of the first frame

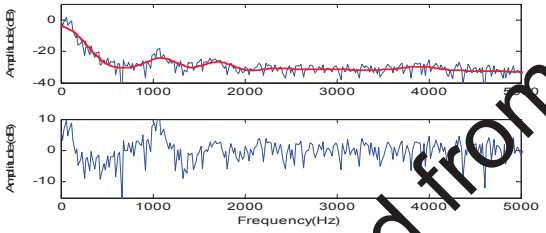


(ii) EMD-based separation of the second frame

(a)



(i) Wavelet-based separation of the first frame



(ii) Wavelet-based separation of the second frame

(b)

Fig. 5: Comparison between wavelet-based and EMD-based separation methods applied to two successive voiced frames extracted from a speech signal corresponding to a sentence uttered by a male speaker for a frame length of 256 samples. (a) EMD-based separation. (b) Wavelet-based separation.

5. CONCLUSION

In this presentation, the empirical mode decomposition algorithm has been proposed as an alternative to decompose the log magnitude spectrum of the speech signal into spectral envelope and harmonic component and its performance has been compared to that of the wavelet-based approach. The proposed method is simple and systematic compared to the wavelet-based method. The results show that the proposed method provides accurate estimates of the harmonic component and spectral envelope for short as well as for long frames.

6. REFERENCES

[1] J. H. Deller, J. G. Proakis, J. H. L. Hansen, Discrete-time processing of speech signals, Prentice Hall, 1993.

[2] M. O. Rosa, R. M., J. C. Pasiral and M. Grellet, "Adaptive estimation of residue signal for voice pathology diagnosis", *IEEE Trans. Biomed. Eng*., vol 47, no 1, pp. 96-104, 2000.

[3] A. V. Oppenheim and R. W. Schafer, Digital signal processing, NJ: Prentice Hall, 1975.

[4] de Krom, G., "A Cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", *J. Speech and Hearing Res.*, vol 36, pp. 254-266, 1993.

[5] M. Noll, "Short-time Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection", *J. Acoust. Soc. Amer*., vol. 36 No. 2, pp. 296-302, 1964.

[6] P. Veprek and M. S. Scordilis, "Analysis, enhancement and evaluation of five pitch determination algorithms", *Speech Comm*., vol. 37, pp. 249:270, 2002.

[7] H. Weiping and R. Linggard, "Speech Signal Deconvolution Using wavelets Filter Banks", WAA '01, Lecture Notes in Computer Science (LNCS), Y. Y. Tang et al. (Eds.), LNCS 2552, pp. 248-256, 2001.

[8] Huang N.E. et al., "The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. R. Soc. London Ser. A* vol. 454, pp. 903-995, 1998.

[9] P. Flandrin, G. Rilling, and P. Goncalves "Empirical mode decomposition as a filter bank", *IEEE Sig. Proc. Lett*., vol. 11, No 2, pp. 112–114, 2004.

[10] Navin Chatlani and John J. Soraghan, "EMD-Based Filtering (EMDF) of Low-Frequency Noise for Speech Enhancement", *IEEE Trans. audio, speech, and lang. proc*., vol. 20, No. 4, pp. 1158-1166, 2012.

[11] L. R. Rabiner,''Digital-Formant Synthesizer for Speech-Synthesis Studies'' , *J. Acoust. Soc. Amer*., vol. 43, No. 4, pp. 822-828, 1968.