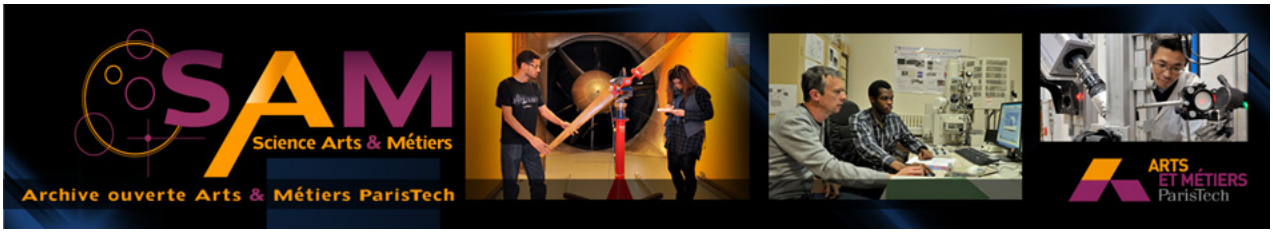
Speech enhancement using empirical mode decomposition and the Teager–Kaiser energy operator

Kais Khaldi, Abdel-Ouahab Boudraa, Ali Komaty



Speech enhancement using empirical mode decomposition and the Teager–Kaiser energy operator

Kais Khaldi
Ecole Nationale d'Ingénieurs de Tunis, Boite Postale 37, Le Belvédère, 1002 Tunis, Tunisia

Abdel-Ouahab Boudraa[a] and Ali Komaty
Ecole Navale, CC 600, 29240 Brest Cedex 9, France

(Received 18 February 2013; revised 12 November 2013; accepted 19 November 2013)

In this paper a speech denoising strategy based on time adaptive thresholding of intrinsic modes functions (IMFs) of the signal, extracted by empirical mode decomposition (EMD), is introduced. The denoised signal is reconstructed by the superposition of its adaptive thresholded IMFs. Adaptive thresholds are estimated using the Teager–Kaiser energy operator (TKEO) of signal IMFs. More precisely, TKEO identifies the type of frame by expanding differences between speech and non-speech frames in each IMF. Based on the EMD, the proposed speech denoising scheme is a fully data-driven approach. The method is tested on speech signals with different noise levels and the results are compared to EMD-shrinkage and wavelet transform (WT) coupled with TKEO. Speech enhancement performance is evaluated using output signal to noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) measure. Based on the analyzed speech signals, the proposed enhancement scheme performs better than WT-TKEO and EMD-shrinkage approaches in terms of output SNR and PESQ. The noise is greatly reduced using time-adaptive thresholding than universal thresholding. The study is limited to signals corrupted by additive white Gaussian noise.
[http://dx.doi.org/10.1121/1.4837835]

I. INTRODUCTION

Degradation of signals by noise is an omnipresent problem. Thus, removal of noise is a key problem in almost all fields of signal processing. In speech, the presence of background noise reduces its quality and intelligibility. Speech enhancement techniques are used to cancel background noises in order to improve the perceptual quality and intelligibility of speech. Many applications such as hearing aids, automatic speech recognition, and voice communication drive the effort to develop more effective noise reduction algorithms for better performance.[1–6] Speech enhancement can be viewed as an estimation problem, where an unknown clean speech signal is to be estimated from its noisy version. This challenging problem is difficult due to the random nature of the background noise and the inherent complexity of the speech. A variety of noise reduction techniques with varying complexity have been developed mostly based on model-based methods, transform domain approaches, and adaptive filtering.[2] Linear methods such as Wiener filtering are largely used because they are easy to implement and to design.[7] However, these methods are not effective when the noise level is unknown or difficult to estimate. To overcome these difficulties, nonlinear methods have been proposed, and especially those based on wavelet thresholding.[8,9] The idea of wavelet thresholding relies on the assumption that signal magnitudes dominate those of the noise in a wavelet representation, so that wavelet coefficients can be set to zero if their magnitudes are less than a pre-determined threshold.[9] However, applying thresholding uniformly to all wavelet coefficients not only cancel the noise, but also some speech components. To solve this problem, a method based on adaptive thresholding has been proposed[10] but a limitation of the wavelet approach is that the basic functions are fixed, and thus do not necessarily match all real signals. Utilizing inappropriate wavelet decomposition will limit the performance of the wavelet-based speech enhancement scheme.

Recently, a new temporal signal decomposition method, called empirical mode decomposition (EMD), has been introduced by Huang *et al.*[11] for analyzing non-stationary and nonlinear time series. This new adaptive expansion decomposes a signal into oscillatory components called intrinsic mode functions (IMFs). These modes are zero-mean with symmetric envelopes AM-FM components. The major advantage of the EMD is that the basic functions are derived from the signal itself. Hence, the analysis is adaptive in contrast to the traditional methods where the basis functions are fixed. We have recently shown that uniform or time-constant thresholding of IMFs followed by their superposition improve the speech denoising results compared to those of the wavelet approach.[12,13] In the present work we improve the performances of this strategy[12] by combining EMD with time adaptive thresholding of the modes. Interesting results are obtained using this approach,[12] in some cases, such as in slight noises contamination, this method suffers from the time-constant thresholding problem. Due to the time variability and non-stationarity of speech signal, a constant threshold can induce an over thresholding of some segments of its extracted modes. In other words an identical threshold will not only suppress unwanted noise but also segments liked unvoiced ones.[10] In addition, this

[a]Author to whom correspondence should be addressed. Electronic mail: boudra@ecole-navale.fr

thresholding does not take into account the waveform of the speech. In this paper, unlike the time-constant threshold based-method, the threshold value is adapted to the waveform of each mode in order to achieve better speech enhancement. More particularly, a frame based approach is used where the calculated threshold values are adapted for speech segments and kept unchanged for non-speech ones. To identify the type of frame we use the Teager–Kaiser energy operator (TKEO) to expand differences between speech and non-speech frames. More precisely, the TKEO provides an estimate of the signal energy over the time, and thus can be used to obtain an estimate of the speech/non-speech activity and then decide an appropriate time adaptive threshold in the speech/non-speech frame. It has been shown that the combination of TKEO with wavelet transform,[10] conventional wavelet packet transform,[14] or perceptual wavelet packet transform[15] is useful for speech enhancement. Unlike the approaches developed in Refs. 10, 14, and 15, in this paper TKEO is combined with the EMD which is a type of adaptive wavelet decomposition well suited for large classes of signals. The basic idea of the new scheme is to pre-process the IMFs using time adaptive thresholding followed by the superposition of their thresholded versions. The proposed method is applied to speech signals corrupted with different noise levels, and the results are compared to the EMD-shrinkage[13] and wavelet transform (WT) using TKEO.[10]

The paper is organized as follows. Section II explains the basics of the EMD and the TKEO. The principle of the denoising is detailed in Sec. III. Results are presented in Sec. IV and conclusions are drawn in Sec. V.

II. EMD-TKEO APPROACH

It is well known that behavior of the cochlea operates as a bank of non-linear dynamic filters. Thus, most speech enhancement approaches use filter banks. In this work we exploit the filter bank aspect of the EMD which deals with nonlinearity and nonstationarity of the speech signal. We have recently shown the interest of the combination of EMD and TKEO for time-frequency analysis and particularly for AM-FM signal demodulation,[16–18] and for linear FM signal detection.[19] We show in this work how this combination can be exploited, in the time domain, for speech enhancement purposes. The TKEO and the EMD are applied to develop a time-adaptive thresholding and an adaptive decomposition for speech enhancement, respectively.

A. EMD basics

The EMD expands any signal $\mathtt{x}(t)$ into a finite number of IMFs through an iterative process called *sifting*; each one with a distinct time scale.[11] The decomposition is based on the local time scale of $\mathtt{x}(t)$, and yields adaptive basis functions. The EMD can be seen as a type of wavelet decomposition whose subbands are built up progressively to separate the different components of $\mathtt{x}(t)$. Each IMF replaces the signal details, at a certain scale or frequency band. The EMD picks out the highest frequency oscillation that remains in $\mathtt{x}(t)$ and thus, locally each mode contains lower frequency oscillations than the one extracted just before. By definition, an IMF satisfies two conditions.

(1) The number of extrema and the number of zero crossings may differ by no more than 1.
(2) The average value of the envelope defined by the local maxima, and the envelope defined by the local minima, is zero.

Thus, locally each IMF contains lower frequency oscillations than the just extracted one. The EMD does not use a pre-determined filter or a wavelet function, and it is a fully data-driven method.[11] To be successfully decomposed into IMFs, the signal $\mathtt{x}(t)$ must have at least two extrema, one minimum and one maximum. An IMF is extracted using an iterative process (sifting) described as follows.[11]

Step 1: Fix the threshold ϵ and set $j \leftarrow 1$ (jth IMF).

Step 2: $\mathtt{r}_{j-1}(t) \leftarrow \mathtt{x}(t)$ (residual), $i \leftarrow 1$ (i number of sifts).

Step 3: Extract the jth IMF,

(a) (a) $\mathtt{h}_{j,i-1}(t) \leftarrow \mathtt{r}_{j-1}(t)$,
(b) extract local maxima/minima of $\mathtt{h}_{j,i-1}(t)$,
(c) compute upper and lower envelopes $\mathtt{U}_{j,i-1}(t)$ and $\mathtt{L}_{j,i-1}(t)$ by interpolating, using cubic spline, respectively, local maxima and minima of $\mathtt{h}_{j,i-1}(t)$,
(d) compute the mean of the envelopes $\mu_{j,i-1}(t) = (\mathtt{U}_{j,i-1}(t) + \mathtt{L}_{j,i-1}(t))/2$,
(e) update $\mathtt{h}_{j,i}(t) \leftarrow \mathtt{h}_{j,i-1}(t) - \mu_{j,i-1}(t)$, $i \leftarrow i+1$,
(f) calculate the stopping criterion $\sigma(i) = \sum_{t=1}^{T} |\mathtt{h}_{j,i-1}(t) - \mathtt{h}_{j,i}(t)|^2 / (\mathtt{h}_{j,i-1}(t))^2$,
(g) repeat steps (b)–(f) until $\sigma(i) < \epsilon$ and then put $\mathtt{IMF}_j(t)$ $\mathtt{h}_{j,i}(t)$ (jth IMF).

Step 4: Update residual $\mathtt{r}_j(t) \leftarrow \mathtt{r}_{j-1}(t) - \mathtt{IMF}_j(t)$.

Step 5: Repeat step 3 with $j \leftarrow j+1$ until the number of extrema in $\mathtt{r}_j(t)$ is ≤ 2.

T is $\mathtt{x}(t)$ time duration and $\leftarrow$ is the assignment operator. The sifting is repeated several times (i) in order to get $\mathtt{h}$ true IMF that fulfills the conditions (1) and (2). At the end of the sifting process the signal $\mathtt{x}(t)$ can be expressed as follows:

$$\mathtt{x}(t) = \sum_{j=1}^{C} \mathtt{IMF}_j(t) + \mathtt{r}_C(t), \tag{1}$$

where C, determined automatically using σ [step 3(f)], is the total number of sifted IMFs and $\mathtt{r}_C(t)$ is the final residue. To guarantee that IMF components retain enough physical sense of both amplitude and frequency modulation, we have to determine the value σ for the sifting. This is accomplished by limiting the size of the standard deviation σ computed from the two consecutive sifting results. Usually, σ (or ϵ) is set between 0.2 to 0.3.[11]

B. Teager–Kaiser energy operator

TKEO is a nonlinear energy tracking operator and its output to a given signal, $\mathtt{x}(t)$, is the actual physical energy required to produce $\mathtt{x}(t)$.[20] An important aspect of this quadratic operator is that it is nearly instantaneous given that, in its discrete version, only three samples are required in the

energy computation at each time instant. This operator amplifies the discontinuities and sudden amplitude changes in $\mathrm{x}(t)$ while the soft transitions between samples are reduced. TKEO has many applications particularly in speech processing.[21,22] Significant research on the theory and applications of the TKEO has been conducted. The majority of the analysis in signal and image processing has mainly dealt with the properties of TKEO-based demodulation algorithms and not with the operator itself.[23,24] In the present work we exploit the output of the TKEO. In continuous time, the TKEO is defined as

$$\Psi[\mathrm{x}(t)] = \left(\frac{d\mathrm{x}(t)}{dt}\right)^2 - \mathrm{x}(t)\frac{d^2\mathrm{x}(t)}{dt^2}. \tag{2}$$

For discrete time signal $\mathrm{x}(n)$, TKEO can be approximated as follows:[21,24]

$$\Psi[\mathrm{x}(\mathrm{n})] = x^2(n) - \mathrm{x}(n+1)\mathrm{x}(n-1). \tag{3}$$

Equation (3) shows that TKEO computes a running estimate of the signal energy at each instant that takes into account the signal strengths at is immediate neighbors. Thus, it is justified that TKEO can be applied to problem of detection as it computes instantaneous energy relative to its immediate neighbors. Also, it is evident from Eq. (3) that TKEO is expected to suppress slowly varying parts of a signal and highlight abruptly changing parts (sudden bursts) of this signal.

III. DENOISING PRINCIPLE

Assuming additive noise, any observed speech signal, $\mathrm{y}(t)$, can be represented as the sum of the clean speech signal $\mathrm{x}(t)$ and the noise $\mathrm{b}(t)$,

$$\mathrm{y}(t) = \mathrm{x}(t) + \mathrm{b}(t). \tag{4}$$

The noisy signal is first decomposed into IMFs and a residual followed by the application of the TKEO to each mode as follows:

$$\mathrm{E}_k(t) = \Psi[\mathrm{IMF}_k(t)]. \tag{5}$$

The idea of the TKEO is to enhance the ability to discriminate speech samples from those of noise. This operator has interesting noise reduction capability.[25] More particularly, the TKEO is expected to expand the difference between the approximation and detail parts of the signal. Detail parts can be attributed to noise. The output of TKEO, $\mathrm{E}_k(t)$, of $\mathrm{IMF}_k(t)$ is smoothed, using a filter, in order to create a mask, $\mathrm{M}_k(t)$, where it is easy to distinguish between speech like-regions and noise-like ones. The normalized version of the TKEO output is given by

$$\mathrm{M}_k(t) = \frac{|\mathrm{h}(t) * \mathrm{E}_k(t)|}{\max(|\mathrm{h}(t) * \mathrm{E}_k(t)|)}, \tag{6}$$

where $*$ is convolution operation, $\mathrm{h}(t)$ is the impulse response of the filter, and "max" is the maximum of the smoothed samples of the TKEO output signal in the considered mode. We use here a second order IIR low-pass filter.[10] In general the obtained mask $\mathrm{M}_k(t)$ presents peaks and valleys where speech dominance is characterized by significant contrast between peaks and valleys, while its absence is characterized by a weaker contrast. To distinguish between these regions (frames) an offset parameter, S_k, that estimates the valley's level is defined.[14] This offset is estimated over the analyzed frame as the abscissa of the maximum of the amplitude distribution, A, of the corresponding mask $\mathrm{M}_k(t)$:

$$\mathrm{S}_k = \mathtt{abscissa}[\max(\mathrm{A}(\mathrm{M}_k(t)))]. \tag{7}$$

Noise dominance is observed with a mask, $\mathrm{M}_k(t)$, of high mean value and with a weaker contrast between peaks and valleys. Thus, the variance of $\mathrm{M}_k(t)$ is close to zero and the corresponding distribution $\mathrm{A}(\mathrm{M}_k(t))$ gives rise to a relevant mode that approaches 1. Consequently, S_k is close to 1. On the contrary, speech dominance is interpreted as a mask, $\mathrm{M}_k(t)$, of low mean value and with a significant contrast between peaks and valleys of $\mathrm{M}_k(t)$ due essentially to the time-variability of the signal envelope. Accordingly, the variance of the mask is more important and the S_k is close to zero.

A. Time invariant threshold

It has been shown that for removing additive white Gaussian noise of an IMF a time-constant threshold can be used.[12,13] This fixed mode-dependent threshold is expressed as follows:[8,12,13]

$$\lambda_k = \tilde{\sigma}_k\sqrt{2\log(T)}, \tag{8}$$

where T is the signal length and the noise level $\tilde{\sigma}_k$ of the kth mode is calculated as the median of the absolute value of the samples of this mode,[12,13]

$$\tilde{\sigma}_k = 1.4826 \times \mathtt{median}\{|\mathrm{IMF}_k(t) - \mathtt{median}\{\mathrm{IMF}_k(t)\}|\}. \tag{9}$$

The constant factor $\tilde{\sigma}_k = 1.4826$ is the 75th percentile of the normal distribution with unit variance. This fixed mode threshold strategy is extremely simple since it does not depend on the extracted mode, but on the noise variance $\tilde{\sigma}_k{}^2$, and works well when noise dominates the observed data.[12,13] However, this thresholding does not perform well when the underlying signal dominates the observed data. To overcome this problem, a time-adapting threshold is necessary.

B. Time-adapting threshold

Due to the time variability and non-stationarity of the speech signal, the time invariant threshold, λ_k, can induce an over-thresholding and thus the perceptual quality of the enhanced signal is degraded.[12] This drawback will be serious when the speech signal is contaminated by slight noises. To overcome this problem, the thresholding is adapted to the speech signal waveform but regardless of its time energy evolution. More precisely, the mask $\mathrm{M}_k(t)$ is pre-processed

so that the calculated thresholds will be well adapted and less dependent on energy variation of speech waveform.[14] The processed mask, noted $\mathtt{M}_k'(t)$, is obtained by suppressing the calculated offset $\mathtt{S}_k$ followed by a normalization of the obtained mask, before applying a root power function of $1/\gamma_k$,

$$\mathtt{M}_k'(t) = \left[\frac{\mathtt{M}_k(t) - \mathtt{S}_k}{\max(|\,\mathtt{M}_k(t) - \mathtt{S}_k\,|)}\right]^{1/\gamma_k}. \tag{10}$$

The parameter γ_k is used in order to implement a compromise between noise removal and speech distortion.[10,14] The value of this parameter depends upon the noise level of the input signal and is determined experimentally (set to 8 in Refs. 10 and 14). Using the modulated mask $\mathtt{M}_k'(t)$ the time adapted threshold, $\mu_k(t)$, is calculated by updating the corresponding λ_k value as follows:[10]

$$\mu_k(t) = \lambda_k(1 - \mathtt{M}_k'(t)). \tag{11}$$

As illustrated by Eq. (11), the time adaptive threshold is made smaller for speech-dominate frame and higher for noisy-like one. Thus, for speech segments, $\mathtt{M}_k'(t)$ is close to 1 and close to 0 for non-speech ones. Note that if $\mathtt{M}_k(t)$ approaches $\mathtt{S}_k$, $\mathtt{M}_k'(t)$ is close to zero and full magnitude of the threshold (that means equal to λ_k) is used. Finally, the threshold is time adapted for only speech-like frames and kept unchanged for noisy-like ones using the calculated offset level $\mathtt{S}_k$ as follows:

$$\tau_k(t) = \begin{cases} \mu_k(t), & \text{if } \mathtt{S}_k \le \zeta_k, \\ \lambda_k, & \text{if } \mathtt{S}_k > \zeta_k, \end{cases} \tag{12}$$

where ζ_k is a parameter value to discriminate speech from silence. As for γ_k parameter, ζ_k is determined experimentally and set to constant value ($\zeta_k = 0.35$).[10,14] Based on simulations, we show that the best parameters γ_k and ζ_k can be obtained by maximizing the output SNR ($\mathrm{SNR}_{\mathrm{out}}$) of the denoised signal (see Fig. 6).

C. Threshold modulation

Finally, using the calculated time adaptive threshold $\tau_k(t)$, a mode $\tilde{f}_k$ can be recovered from its noisy version $\mathtt{IMF}_k$ using different thresholding strategies.[26] The most adapted strategy for speech enhancement is soft shrinkage.[12,13] It is the one used in this paper and is given by

$$\tilde{f}_k(t) = \begin{cases} \mathtt{sgn}(\mathtt{IMF}_k(t))(|\mathtt{IMF}_k(t)| - \tau_k(t)), & \text{if } |\mathtt{IMF}_k(t)| \ge \tau_k(t), \\ 0, & \text{if} |\mathtt{IMF}_k(t)| < \tau_k(t), \end{cases} \tag{13}$$

where τ_k is the threshold of the mode $\mathtt{IMF}_k$ and $\mathtt{sgn}(z)$ is the sign function of z. The denoised speech signal is constructed with the inverse transformation $\mathtt{EMD}^{-1}$ of the thresholded $\mathtt{IMF}$s as follows:

$$\tilde{\mathtt{x}}(t) = \sum_{k=1}^{C} \tilde{f}_k(t) + \mathtt{r}_C(t). \tag{14}$$

The different steps of the proposed denoising scheme are shown in Fig. 1.

IV. RESULTS

The proposed denoising method is tested on eight speech signals sampled at 16 kHz and are shown in Figs. 2 and 3. These figures illustrate the variety of speech contents met in the different files that we analyzed. The signals are corrupted by additive white Gaussian noise, ranging from −10 to 10 dB, whose level is fixed through the input signal to noise ratio (SNR). Obtained results are compared to those of EMD-shrinkage[13] and WT-TKEO.[10] The Daubechies wavelet Db8 is used as mother wavelet. The corresponding basis is orthonormal, ensuring that the decomposed signal is reconstructed without the presence of residues due to asymmetries of the wavelet mother function. These features make Db8 wavelet a good candidate as denoising tool. As an objective criterion to evaluate the performance of the denoising method, we use the perceptual evaluation of speech quality[27] (PESQ) and the output SNR. PESQ values range from 4.5 (the highest quality of speech) down to −0.5 and are known for their high correlation with subjective quality scores. The PESQ score is adopted as an objective measure tool for predicting the overall quality of enhanced speech. Also, to evaluate the subjective observation of enhanced speech, spectrograms of the clean speech, the noisy speech and the enhanced speech signal s obtained using the different denoising methods are presented. Figure 4(a) shows the first extracted mode, $\mathtt{IMF}_1(t)$, of speech #1 signal. Note that this high frequency mode is very noisy. Figures 4(b) and 4(c) illustrate the waveforms of the associated output of TKEO, $\mathtt{E}_1(t)$, and its corresponding mask, $\mathtt{M}_1(t)$, respectively. Figure 5 shows the standard threshold $\mu_1(t)$ and the time adaptive threshold λ_1 of this mode. Unlike the standard threshold, the adaptive threshold value is changing over time and it is dependent on the speech-like regions. Parameters γ_k and ζ_k [Eqs. (10) and (12)] in general are determined experimentally ($\zeta_k = 0.35$, $\gamma_k = 8$).[10,14] To the best of our knowledge, up to now there are no strategies to automatically find such parameters. Based on simulations, we show in paper that the couple of these parameters, maximizes $\mathrm{SNR}_{\mathrm{out}}$ of the processed signals. $\mathrm{SNR}_{\mathrm{out}}$ as a function of ζ_1 and γ_1 of $\mathtt{IMF}_1(t)$ at $\mathrm{SNR}_{\mathrm{in}} = -5$ dB is depicted in Fig. 6. This figure

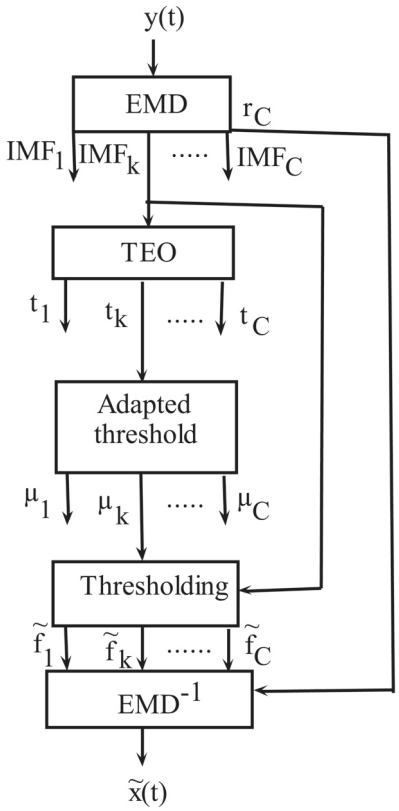


FIG. 1. Denoising scheme.

reveals a peak of SNR_{out} of 11 dB at $(\zeta_1 = 0.37, \gamma_1 = 8)$ and this result confirms the findings of Bahoura and Rouat.[10,14] These two optimal values with $\mu_1(t)$ and λ_1 [Eq. (12)] are used for the denoising of $\mathtt{IMF}_1(t)$. The thresholded mode, $\tilde{f}_1(t)$, is shown in Fig. 4(d). It is easy to see from this figure that $\mathtt{IMF}_1(t)$ is enhanced using time-adapting threshold. Denoised versions of speech #1 signal obtained by the proposed approach, WT-TKEO (Ref. 10) and EMD-shrinkage,[13] are shown in Fig. 7. A careful comparative examination of signals of Figs. 7(a)–7(d) shows that the proposed method performs better than the other approaches in terms of noise reduction. The signal is well reconstructed and its features

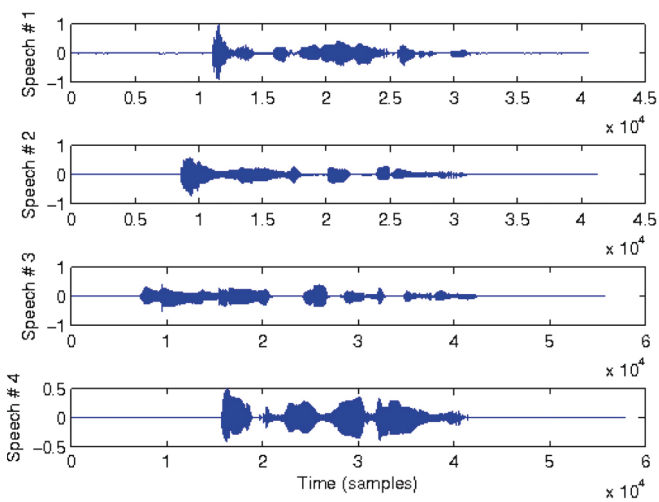


FIG. 2. (Color online) Original signals (1,2,3,4).

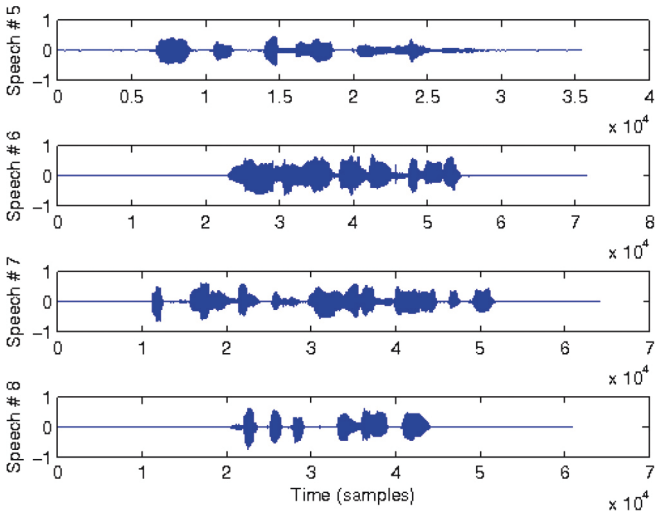


FIG. 3. (Color online) Original signals (5,6,7,8).

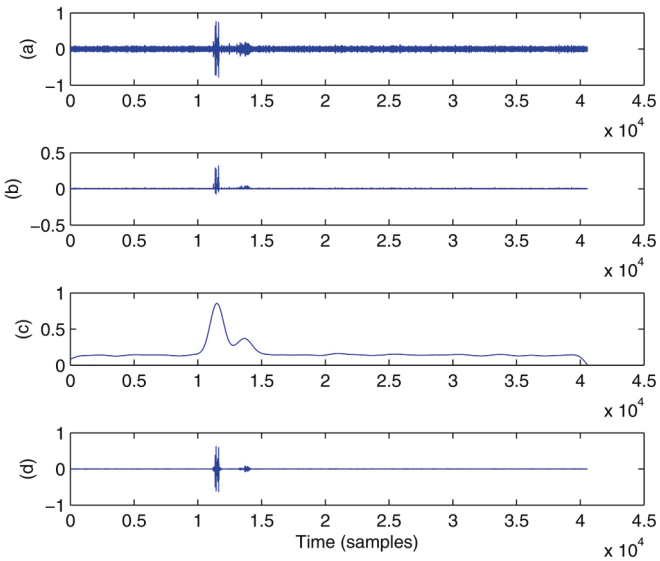


FIG. 4. (Color online) (a) $\mathtt{IMF}_1(t)$ of signal #1. (b) $\mathtt{E}_1(t)$. (c) $\mathtt{M}_1(t)$. (d) $\tilde{f}_1(t)$.

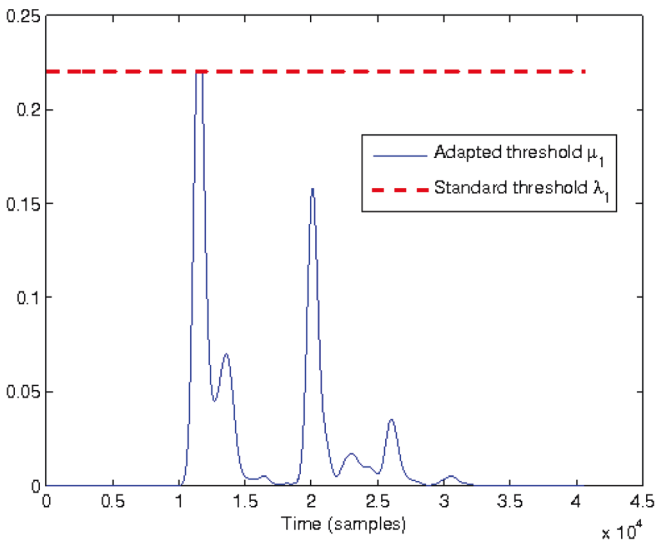


FIG. 5. (Color online) Standard and adapted thresholds of $\mathtt{IMF}_1$ of signal #1.

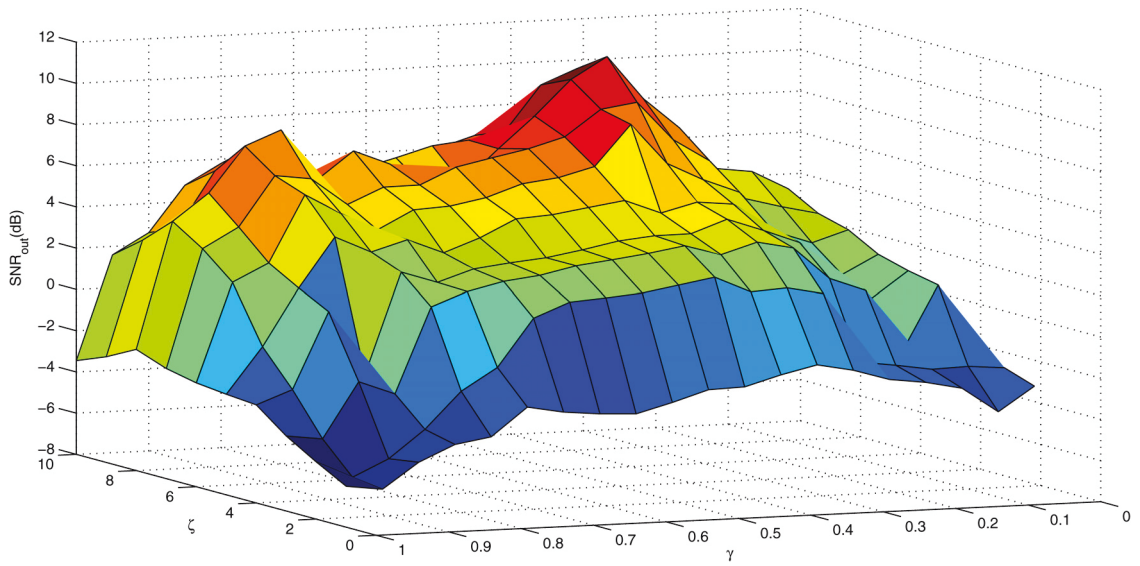


FIG. 6. (Color online) SNR_{out} versus ζ_1 and γ_1 for speech 1 signal at $SNR_{in} = -5$ dB.

(details) well preserved. The spectrograms of clean, noisy and processed signals are illustrated in Figs. 7(f)–7(j). It is clear from these figures that the amount of distortion is greatly reduced and most parts of the speech are preserved by the proposed approach [Fig. 7(j)] in comparison with the other two methods [Figs. 7(h), 7(i)]. Specifically, both WT-TKEO (Ref. 10) and EMD-shrinkage[13] lead to visible residual noise and noticeable speech distortion. Thus, the spectrogram observations with lower distortion validate our claim of better speech signal reconstruction by our approach in comparison to the two other methods. This fact is confirmed by the results reported in Fig. 8 (speech #1, speech #2, and speech #3) of gain SNR_{out} values achieved by the proposed method compared to other methods. The results are averaged over 100 noise realizations. For deeper performance investigation, Fig. 8 reports the variations of the SNR_{out} versus the SNR_{in} corresponding to the denoising of the three speech signals. These results demonstrate the effectiveness of the proposed method across the SNR_{in} values. The achieved SNR_{out} improvement is much higher than those obtained by WT-TKEO and EMD-shrinkage. For $SNR_{in} \leq 15$ dB, the proposed method outperforms WT-TKEO and EMD-shrinkage which highlights the interest to combine EMD and TKEO as a basis of an adaptive denoising strategy. Under higher SNR conditions ($SNR_{in} > 15$ dB) where speech signal is slightly contaminated, the proposed scheme and WT-TKEO performs similarly and avoid the over thresholding of EMD-shrinkage. SNR_{out} gains obtained by the proposed technique varies from 5.2 to 19 dB and, in particular, even for very low SNR_{in} values, we can still observe the effectiveness of the proposed method in removing the noise components. Note that, overall, WT-TKEO performs better than EMD-shrinkage and this is mostly attributed to the time-adaptive thresholding strategy. But, the superiority of the

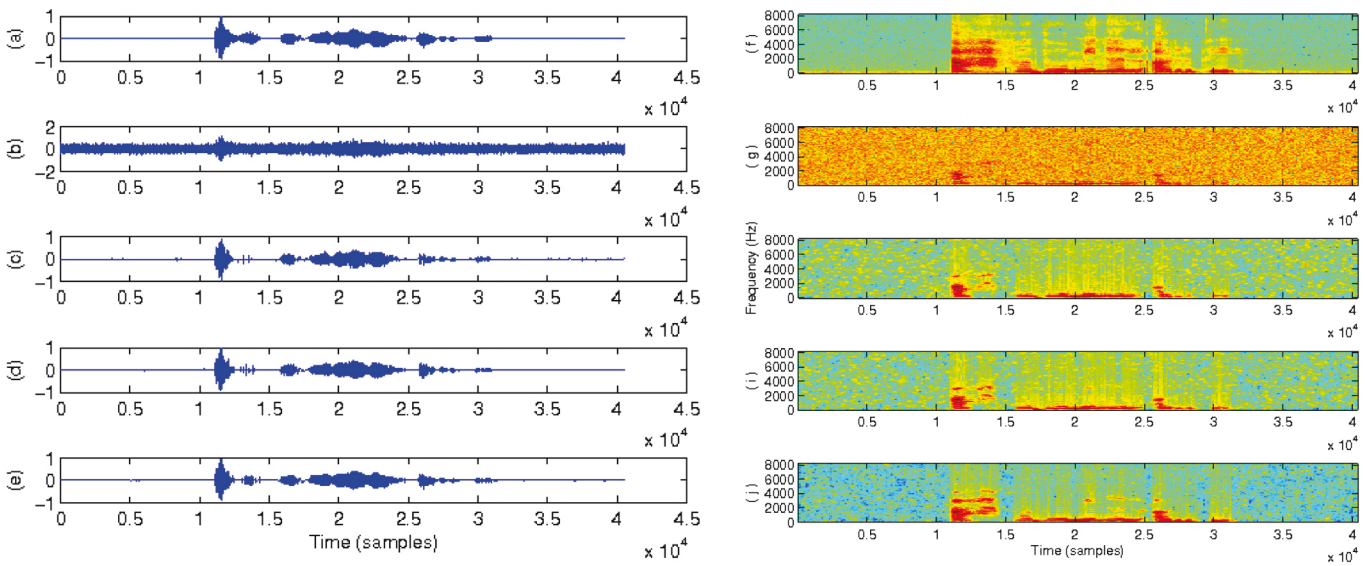


FIG. 7. (Color online) Speech denoising results: (a) original signal, (b) noisy signal (SNR = −5 dB), (c) EMD-shrinkage (Ref. 13), (d) WT-TKEO (Ref. 10), (e) proposed approach, and the associated spectrograms depicted respectively in (f)–(j).

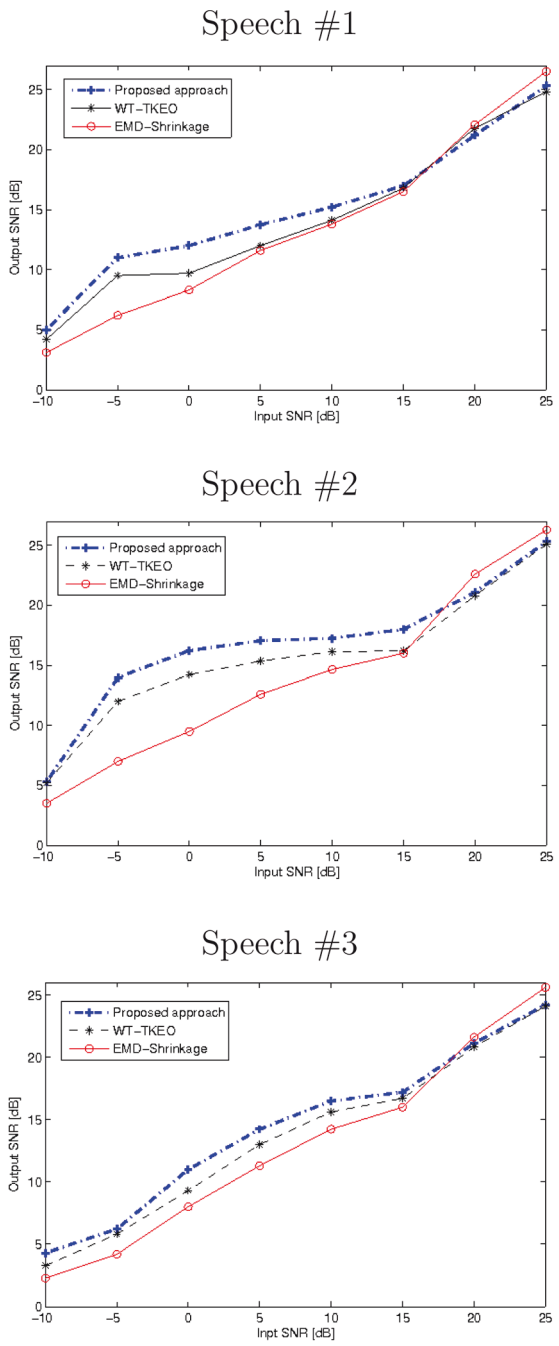


FIG. 8. (Color online) Variations of SNR_{out} versus SNR_{in} for three signals. The reported results correspond to the EMD-shrinkage (Ref. 13), WT-TKEO (Ref. 10), and the proposed approach.

proposed method over WT-TKEO is essentially due to the adaptive decomposition of the speech signal provided by EMD. Figure 9 shows the PESQ scores of the enhanced three speech signals obtained from our proposed method in the same plot with those of WT-TKEO (Ref. 10) and EMD-shrinkage.[13] It is clear that the proposed method produces the best PESQ scores in both low and high SNR regions. These high scores are indicative of the good speech quality obtained with the proposed method. The best result is obtained for speech #3 for SNR ≥ 5 dB. In addition to the first three signals, the method has also been tested on other speech signals whose performance is reported in Table I. It

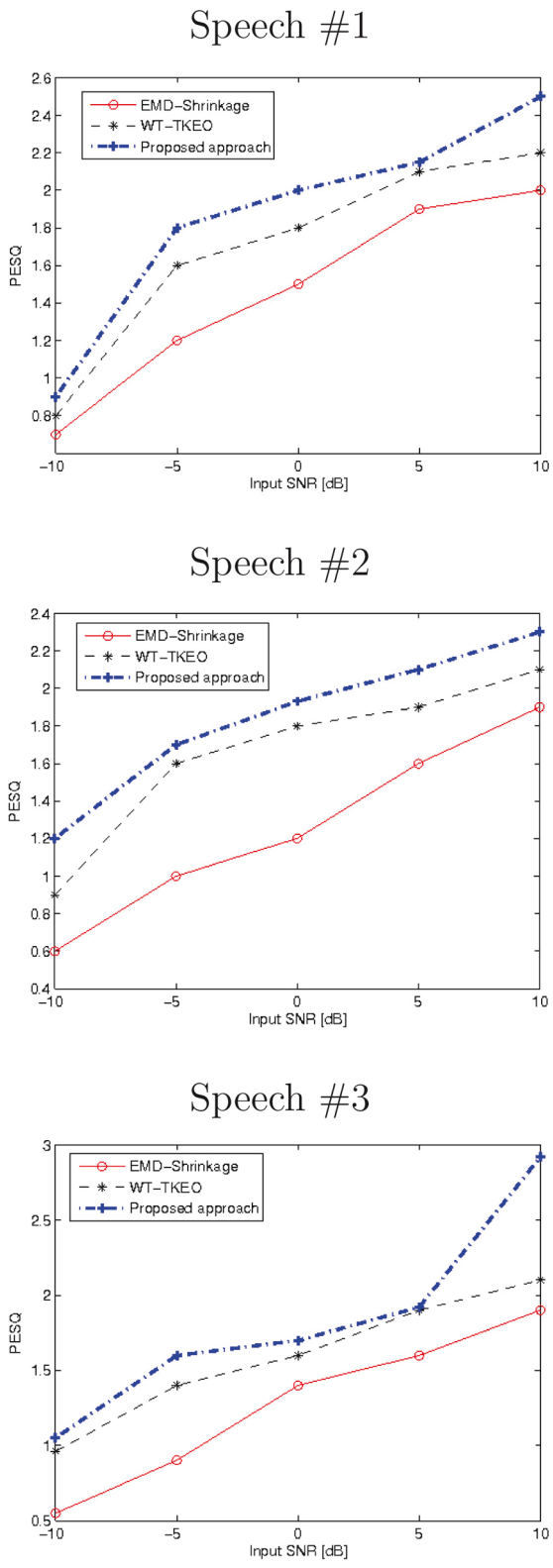


FIG. 9. (Color online) Variations of PESQ values versus SNR_{in} for three signals. The reported results correspond to the EMD-shrinkage (Ref. 13), WT-TKEO (Ref. 10), and the proposed approach.

follows from this table that the proposed method performs better than WT-TKEO (Ref. 10) and EMD-shrinkage[13] in terms of SNR_{out} and PESQ scores.

The analysis of the obtained results emphasizes the usefulness of the time-adaptive thresholding combined with

TABLE I. Denoising results of eight speech signals by the proposed approach, EMD-shrinkage and WT-TKEO at $SNR_{in} = 5$ dB.

	Signals	#1	#2	#3	#4	#5	#6	#7	#8
EMD-shrinkage	SNR_{out}	10.7	12	10.3	11	11.8	10.7	12	11.1
	PESQ	1.17	1.02	0.81	0.93	1.13	0.82	1.2	1.03
WT-TKEO	SNR_{out}	11.2	14.3	13	13.6	12.4	13.2	13.5	12.2
	PESQ	2	1.85	1.81	1.93	1.52	1.76	1.71	1.36
Proposed approach	SNR_{out}	14	16.8	14.5	15.2	14.6	15.8	13.6	13.3
	PESQ	2.1	2.05	1.82	2.02	1.91	2.06	1.85	1.54

adaptive decomposition provided by the EMD as a filter bank. These results also confirm our findings on the interest of EMD as filter bank for speech processing.[12,13] The noise is greatly reduced using time-adaptive thresholding compared to universal thresholding. Instead of EMD we have also tested the ensemble EMD (EEMD) which avoids the mode mixing of conventional EMD. Based on the signals analyzed and from the obtained results we have noted no improvement or differences compared to those given by the conventional EMD. Since EMD is well dedicated to process both stationary and non-stationary signals, our denoising scheme can be applied to a large class of signals (Biomedical, etc.). Further, unlike the approaches developed in Refs. 10 and 15, where the level of the decomposition is fixed empirically, in the proposed method this level is a data-driven parameter. Performance of the denoising approach depends on the quality of the sifting which in turns depends on the way interpolation of the envelopes is performed. Thus, utilizing an inappropriate interpolating function will limit the performance of the EMD-based coding scheme. Recent study has shown that trigonometric interpolation is useful from an analytic point of view, but computationally it is much more expensive than splines. The authors of this study do not recommended it instead of splines interpolation.[28] Thus, in this paper, B-splines are used for interpolation of the modes. Denoising results of the EMD-TKEO are not prejudiced by a pre-determined basis and/or subband filtering process. However, this approach requires the optimization of the selection of parameters (ζ, γ) and the fixed second order IIR filter, $\mathrm{h}(t)$, to construct the mask [Eq. (6)]. For further speech enhancement quality improvement, the impulse response $\mathrm{h}(t)$ must be adapted to each IMF.

V. CONCLUSION

An improved EMD-based approach to speech enhancement is presented. By combining the adaptive nature of EMD and TKEO, the proposed scheme avoids the over-thresholding of segments especially when the speech is just contaminated by slight noise. The TKEO is applied to each IMF to enhance the discriminability of speech and non-speech frames. The discriminatory threshold is time-adapted to the speech waveform, unlike the time-constant universal thresholding. Based on EMD and TKEO, the denoising approach is adaptive and easy to implement. Obtained results for clean speech signals corrupted with additive Gaussian noise with different SNR values ranging from -10 to 10 dB show that the proposed method, associated with the time-adapting threshold, performs better than EMD-shrinkage and WT-TKEO in terms of output SNR and PESQ scores. These results show that the proposed approach is effective for noise removal. The obtained results also show the efficiency of the time adapted threshold to the different components (IMFs) of the signal instead of the standard threshold. To confirm the obtained results and the effectiveness of the proposed approach, the scheme must be evaluated with a large class of speech signals and in different experimental conditions such as different sampling rates, multiplicative noise, or the type of noise. Also as future work we plan to find a strategy to optimize the selection of parameters ζ and γ.

ACKNOWLEDGMENT

The authors would like to thank Professor Mohamed Bahoura from Université de Québec à Rimouski for fruitful discussions on time adaptive thresholding.

[1]S. Vaseghi, *Multimedia Signal Processing: Theory and Applications in Speech, Music and Communications* (Wiley, Chichester, 2007).
[2]P. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, 2007).
[3]T. van den Bogaert, S. Doclo, J. Wouters, and M. Moonen, "Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids," J. Acoust. Soc. Am. **125**, 360–371 (2009).
[4]G. Kim, Y. Lu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494 (2009).
[5]G. Kim and P. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," J. Acoust. Soc. Am. **130**, 1581–1596 (2011).
[6]S. Srinivasan and D. H. R. Naidu, "Speech enhancement using a generic noise codebook," J. Acoust. Soc. Am. **132**, EL161–EL167 (2012).
[7]J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications* (Prentice-Hall, Upper Saddle River, NJ, 1996).
[8]D. Donoho and I. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," Biometrica **81**, 425–455 (1994).
[9]D. Donoho, "De-noising by soft-thresholding," IEEE Trans. Inform. Theory **41**, 613–627 (1995).
[10]M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," IEEE Signal Process. Lett. **8**, 10–12 (2001).
[11]N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," Proc. R. Soc. London **454**, 903–995 (1998).
[12]K. Khaldi, A. O. Boudraa, A. Bouchikhi, M. T.-H. Alouane, and E. S. Diop, "Speech signal noise reduction by EMD," in *Proceedings of IEEE ISCCSP*, Limassal, Cyprus (2008), pp. 1–5.
[13]K. Khaldi, A. O. Boudraa, A. Bouchikhi, and M. T.-H. Alouane, "Speech enhancement via EMD," EURASIP J. Appl. Signal Process. **2008**, 873204 (2008).
[14]M. Bahoura and J. Rouat, "Wavelet speech enhancement based on time-scale adaptation," Speech Commun. **48**, 1620–1637 (2006).
[15]S. Chen and J. Wang, "Speech enhancement using perceptual wavelet packet decomposition and Teager energy operator," J. VLSI Signal Proc. Syst. Signal, Image, Video Technol. **36**, 125–139 (2004).
[16]A. O. Boudraa, J. Cexus, F. Salzenstein, and L. Guillon, "IF estimation using empirical mode decomposition and nonlinear Teager energy operator," in *Proceedings of IEEE ISCCSP*, Hammamet, Tunisia (2004), pp. 45–48.
[17]J. Cexus and A. Boudraa, "Nonstationary signals analysis by Teager–Huang transform (THT)," in *Proceedings of EUSIPCO*, Florence, Italy (2006), pp. 1–5.
[18]A. Bouchikhi and A. O. Boudraa, "Multicomponent AM-FM signals analysis based on EMDB-splines ESA," Signal Process. **92**, 2214–2228 (2012).

[19]J. Cexus, A. Boudraa, and A. Bouchikhi, "A combined Teager–Huang and Hough transforms for LFM signals detection," in *Proceedings of IEEE ISCCSP*, Limassal, Cyprus (2010), pp. 1–5.
[20]J. Kaiser, "Some useful properties of Teager's energy operators," in *Proceedings of ICASSP* (1993), pp. 149–152.
[21]P. Maragos, T. Quatieri, and J. Kaiser, "Speech nonlinearities, modulation and energy operators," in *Proceedings of ICASSP* (1991), pp. 421–424.
[22]F. Jabloun, A. E. Cetin, and E. Erzin, "Teager energy based feature parameters for speech recognition in car noise," IEEE Signal Process. Lett. **6**, 259–261 (1999).
[23]A. P. D. Dimitriadis and P. Maragos, "A comparison of the squared energy and Teager–Kaiser operators for short-term energy estimation in additive noise," IEEE Trans. Signal Process. **57**, 2569–2581 (2009).
[24]A. O. Boudraa, J. Cexus, and K. Abed-Meraim, "Cross-$\Psi_{\mathbb{B}}$-energy operator-based signal detection," J. Acoust. Soc. Am. **123**, 4283–4289 (2008).
[25]H. Patil and S. Viswanath, "Effectiveness of Teager energy operator for epoch detection from speech signals," Int. J. Speech Technol. **14**(4), 321–337 (2011).
[26]S. Mallat, *A Wavelet Tour of Signal Processing* (Academic Press, San Diego, 1999).
[27]ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," *ITU-T recommendation* P.835 (2003).
[28]S. Hwaley, L. E. Atlas, and H. Chizeck, "Some properties of an empirical mode type signal decomposition algorithm," IEEE Signal Process. Lett. **17**, 24–27 (2010).