Audio coding via EMD

Abdel-Ouahab Boudraa, Kais Khaldi, Thierry Chonavel, Mounia Turki Hadj-Alouane, Ali Komaty

Audio coding via EMD

Abdel-Ouahab Boudraa[a], Kais Khaldi[b,d], Thierry Chonavel[c], Mounia Turki Hadj-Alouane[d] and Ali Komaty[e]

[a]*IRENav, Ecole Navale/Arts et Metiers Institute of Technology, CC 600, 29240 Brest Cedex 9, France.*
[b]*College of Science and Arts-Tabarjal, Jouf University, P.O.Box 2014, Al-Jouf, Skaka,42421, KSA.*
[c]*IMT-A, LabSTICC, BP 832, 29285, Brest, France.*
[d]*L3S, ENIT, El-Manar University, BP 37, Le Belvédère, 1002 Tunis, Tunisia.*
[e]*University of Sciences and Arts (USAL) in Lebanon, Beirut, Lebanon.*

Abstract

In this paper an audio coding scheme based on the empirical mode decomposition in association with a psychoacoustic model is presented. The principle of the method consists in breaking down adaptively the audio signal into intrinsic oscillatory components, called Intrinsic Mode Functions (IMFs), that are fully described by their local extrema. These extrema are encoded. The coding is carried out frame by frame and no assumption is made upon the signal to be coded. The number of allocated bits varies from mode to mode and obeys to the coding error inaudibility constraint. Due to the symmetry of an IMF, only the extrema (maxima or minima) of one of its interpolating envelopes are perceptually coded. In addition, to deal with rapidly changing audio signals, a stationarity index is used and when a transient is detected, the frame is split into two overlapping sub-frames. At the decoder side, the IMFs are recovered using the associated coded maxima, and the original signal is reconstructed by IMFs summation. Performance of the proposed coding is analyzed and compared to that of MP3 and AAC codecs, and the wavelet-based coding approach. Based on the analyzed mono audio signals, the obtained results show that the proposed coding scheme outperforms the MP3 and the wavelet-based coding methods and performs slightly better than the AAC codec, showing thus the

*Corresponding author: A.O. Boudraa (boudra@ecole-navale.fr)

potential of the EMD for data-driven audio coding.

Keywords: Empirical mode decomposition, Empirical mode compression, Audio coding, Sub-band coding, Stationarity index, Psychoacoustic model.

---

1. Introduction

Signal coding is a central topic in signal and images processing [1]-[2] and particularly in speech domain where different strategies have been proposed [3]-[4]. In many applications such as digital audio broadcasting or multimedia, low Bit Rate (BR) and high fidelity are required. To reduce the BR, sub-band coding and transform coding approaches have been used to design efficient coding algorithms [5]-[12]. These methods use pre-determined basis functions and perceptual encoding of the significant transform coefficients, following the principle: *"do not code what the ear cannot listen"*. Applying this principle enables good results at low BR. Unfortunately, using fixed basis functions prevents the decomposition from being sparse for a large class of audio signals. As a matter of fact, even if a basis is well suited for a class of audio signals, in the sense that it yields compact descriptions with only a few significant terms, generally there exist other audio signals for which the basis under consideration performs poorly. Thus, there is a need for data driven coding strategies. In [13], an expansion method, referred to as Empirical Mode Decomposition (EMD) has been introduced for analyzing non-stationary data derived from linear or non-linear systems in a totally adaptive way. The EMD has found many applications in audio and speech processing such as audio watermarking [14], speech enhancement [15] and speech classification [16]. The main interest of such decomposition relies on the fact that it involves no prior choice of filters or basis functions. Compared to the classical kernel based approaches, the EMD is a fully data driven technique that recursively breaks down any signal into a reduced set of zero-mean Amplitude Modulated and Frequency Modulated (AM-FM) components called Intrinsic Mode Functions (IMFs). The decomposition starts from finer temporal scales (high frequency IMFs) to coarser ones (low frequency IMFs) where

the IMFs are very well described by their local extrema [13]. Furthermore, the extracted IMFs are almost orthogonal and summing all these IMFs with the residue recovers the original signal, within machine precision [13].

An IMF is an oscillating function that can be fully described by its extrema. So, the IMF can be recovered easily from its extrema by using spline interpolation. Thus, a salient property of the IMF is that it can be fully described by its extrema and representing a signal with these extrema (amplitude and location) can be used for signal and images coding purposes [17],[18]. In particular, this EMD-based strategy can be considered to encode audio signals. The idea is to use signal-adaptive IMFs to replace de facto modified discrete cosine transform (MDCT), that has been used extensively for modern audio codecs. Since the number of extrema decreases from one IMF to the next one, IMFs encodings are not all equally demanding and the number of bits needed to code each IMF will vary, depending whether they represent low or high frequencies of the signal. According to the kind of signal to be processed, the number of extrema to be coded can be reduced using an appropriate thresholding. In this way, for audio signals, extrema can be thresholded using a psychoacoustic model [19] and the IMFs are recovered using these coded extrema. The reduction of the number of extrema encoded, controlled by the perceptual masking curve, yields interesting improvement in compression gain while preserving the listening quality. In addition, in this paper we show that performance can be more improved by taking into account the symmetry of the IMFs and the transient detection in the case of rapidly changing input signals. The main contributions of this paper are described as follows:

- Unlike the approach developed in [19], the symmetry of the envelopes defined by the local maxima and the local minima is exploited. Only the extrema of one envelope are perceptually encoded. At the decoder side, the envelope of maxima (or minima) is reconstructed and the other envelope is deduced by symmetry. Consequently, the BR is approximately

halved.

- For ensuring encoding effectiveness, rapid changes (transients) that may occur inside frames are detected by using a stationarity index. Then, a frame showing sudden change is split into two overlapping sub-frames to reduce the effect of the transition in the associated IMFs.

- Both extrema selection and their quantization are based on a psychoacoustic model. The number of bits used to encode each IMF is adjusted so that the Power Spectral Density (PSD) of the reconstruction error of the IMF remains below the perceptual masking threshold of the audio signal frame.

The proposed coding strategy, referred to as Empirical Mode Compression (EMC), is applied to real mono audio signals, and the results are compared to those of MP3 (ISO/IEC 11172-3 MPEG Layer 3) [20], AAC (ISO/IEC 13818-7 Advanced Audio Coding) [21] codecs and the wavelet-based coding.

2. EMD principle

Unlike standard decomposition methods that project data onto a predefined basis function (harmonic, wavelet), bases of the EMD are derived from the data in a nonlinear and non-stationary way. The EMD decomposes univariate data into slow and fast oscillations. More precisely, the EMD expands, in adaptive way, any real-valued signal $x(t)$ into a limited number of oscillating components, IMFs. Being fully data driven, the IMFs represent the inherent temporal modes (scales) that characterizes $x(t)$. The decomposition yields signal adapted orthogonal basis functions. The EMD can be seen as a type of wavelet decomposition, the sub-bands of which automatically adapt to split the different components of $x(t)$. The IMFs are extracted from $x(t)$ by means of an iterative algorithm called the sifting process [13]. These IMFs are designed to be narrow-band (single-scale). By construction, each IMF is a zero-mean

waveform, number of zero-crossings of which differ at most by one from the number of its extrema. By definition, an IMF satisfies two conditions: (i) the number of extrema and the number of zero crossings may differ by no more than one and (ii) the mean value at each point of the envelope defined by the local maxima, and the envelope defined by the local minima, is zero. Each IMF contains lower frequency oscillations than the ones just extracted before. To be successfully decomposed into IMFs, a signal $x(t)$ of length L_s must have at least two extrema, one minimum and one maximum. The EMD ends up with an expansion of $x(t)$ into IMFs of the form:

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

where C is number of IMFs. The component $\mathtt{r}_C(t)$ is called the residual of the decomposition and cannot contain a full oscillation. It represents the trend within $x(t)$ [13]. The modes $\{\mathtt{IMF}_j(t)\}_{j=1}^{C}$ in equation (1) represent the bases of $x(t)$, and are sparse in the sens that C is much lower than the signal length L_s and template free. To guarantee that the IMFs retain enough physical sense of both amplitude and frequency modulations, a stopping criterion is determined by using a convergence test of Cauchy type. Specifically, the test requires the normalized squared difference between two successive siftings, noted SD (standard deviation), to be small. If SD is smaller than a predetermined value ϵ, usually set between 0.2 and 0.3, the sifting will be stopped [13]. The sifting process of the EMD is summarized in the Algorithm 1:

3. EMD-based coder

In this section, we first describe the structure of the proposed coder. Issues such as the bit allocation, the IMF symmetry and the detection of the non-stationarity of the frame are described in the following subsections.

3.1. Coder principle

The proposed coder is based on the encoding of the IMFs extrema. But, due to the symmetry of the IMFs, only their minima or maxima need to be encoded.

Algorithm 1: Sifting process of EMD, $j \in \{1, 2, \ldots, C\}$.

Input: $x(t)$.

Outputs: $\{\texttt{IMF}_j(t)\}_{j=1}^{C}$, $\texttt{r}_C(t)$.

1. Initialize $x_0(t) = x(t)$.
2. Find the instants of location of all extrema of $x_0(t)$.
3. Interpolate (local spline interpolation) between maxima (minima) to obtain the upper (lower) envelope connecting the maxima, $e_{max}(t)$ ($e_{min}(t)$).
4. Compute the local mean $\text{mean}(t) = (e_{min}(t) + e_{max}(t))/2$.
5. Subtract the local mean from the signal to obtain the oscillatory mode, $d(t) = x_0(t) - \text{mean}(t)$.
6. If $d(t)$ satisfies the stopping criterion, set $\texttt{IMF}_j(t) = d(t)$ else set $x_0(t) := x_0(t) - \texttt{IMF}_j(t)$. If $x_0(t)$ becomes a monotonic function, or does not contain enough extrema to form meaningful envelope, stop the sifting process with $\texttt{r}_C(t) = x_0(t)$. Otherwise, go to step 2.

Without loss of generality, in the rest of this paper we focus on the coding of the maxima.

Maxima coding

Let j and k be the index of the IMF and of the frame respectively. We denote by l the index of a maximum of the IMF and we refer by N_j^k the number of maxima of this IMF. For j^{th} IMF of the k^{th} frame, the l^{th} maximum, $l \in \{1, 2, \ldots, N_j^k\}$, is represented by its amplitude m_{jl}^k and its time index t_{jl}^k (position). Values m_{jl}^k are scaled by a factor equal to $\gamma_j^k = \max_l(m_{jl}^k)$. In practice the sifted IMF is not truly symmetric and thus some amplitude offset, α_j^k, may appear in the j^{th} IMF of the k^{th} frame $\texttt{IMF}_j^k(t)$. Thus we account for it in the encoding process. In

addition, the audio signal is decomposed into overlapping frames of equal length and abrupt changes in signal statistics may occur inside a frame. To account for such events, we possibly split frames into 2 sub-frames.

For bit allocation, the number of bits allocated to the j^{th} IMF of the k^{th} frame is adjusted so that the PSD, $\Gamma_j^k(f)$, of the reconstruction error $\epsilon_j^k(t) = \mathtt{IMF}_j^k(t) - \tilde{\mathtt{IMF}}_j^k(t)$ does not exceed the associated perceptual masking threshold $\mathtt{TM}_k(f)$, where $\tilde{\mathtt{IMF}}_j^k(t)$ is the j^{th} reconstructed IMF of the k^{th} frame (Fig. 6). There is one psychoacoustic evaluation per frame. The encoding process applied frame by frame is summarized as follows:

The EMC scheme

The block diagram of the proposed coding scheme is presented in figure 1. The EMC is given by the following steps.

1. **Segment** the original audio signal into N frames,
2. **Split** the frame into two overlapping sub-frames if a transient is detected,
3. **Extract** using EMD, the modes $\mathtt{IMF}_j^k(t)$ and the residual $\mathtt{r}_C^k(t)$ of the k^{th} frame or sub-frame,
4. **Determine** the values m_{jl}^k, t_{jl}^k, γ_j^k and α_j^k of the j^{th} IMF in the k^{th} frame,
5. **Quantize** and **Encode** these values.

3.2. Windowing process

In speech processing the input signal is usually segmented into quasi-stationary overlapping frames. Many audio codecs operate on frames, and for comparison purpose, the proposed coding is also carried out frame by frame, and no assumption is made on the length of the signal to be coded. For codecs like AAC, smaller frame size is used for transient detection. Indeed for rapidly changing input signals long frames are unfavorable because the temporal spread quantizations will lead to so-called "pre-echoes". However, in practice applying the EMD to frame of very short length can induce end effects problem. Thus, we increase the length of the frame and the detection of the transient is based on

the computation of a stationarity index. When a transient is detected the frame is subdivided into two overlapping sub-frames.

3.3. Transient detection

The concept of transient is not easy to describe precisely [22]. The transient component comes from unwanted measurement artefacts and from instruments that are played very impulsively such as piano, xylophone, or castanet. For example, drums and percussions both have strong transients. More precisely, in the case of acoustic instruments, the transient often corresponds to the period during which the excitation (e.g., a hammer strike) is applied and then damped, leaving only the slow decay at the resonance frequencies of the body [22]. An extensive amount of research has already focused on transient detection [22],[23][24],[25]. To detect the instants of such abrupt changes, spectral properties of the signal can be measured over the time by using a Stationarity Index (SI). Based on the SI tracking, a frame showing sudden change can be split into two adjacent sub-frames. In this work the SI is derived from the representation of a Time-Frequency Distribution (TFD), say $\rho(t, f)$, of the analyzed audio signal [26]. The transient detector does not use prior knowledge of the input signal. As illustrated in figure 2, at each time instant, the index is calculated as a distance between two TFDs, calculated on the right and the left sub-images on both sides of that instant. The measured distance is sensitive to abrupt changes in the spectral characteristics of the signal [26]. The TFD is computed over the audio frame duration. If there is no significant change, the distance remains quite constant, while it shows peaks if a change occurs. For transient signals, the SI has a distinguishable maximum, located where signal change occurred [26]. To measure the difference between two time-shifted sub-images in the TF plane, different distances such as Kullback divergence, Kolgomorov distance, Bhattacharyya distance or "Jensen-like" divergence can be used. In particular, the Bhattacharyya distance is known to be sensitive to abrupt changes of signals in the TF plane [26] and is used as a SI.

Let $\mathtt{I}_1(t; \tau, f)$ and $\mathtt{I}_2(t; \tau, f)$ be two sub-images (Fig. 2), computed at each in-

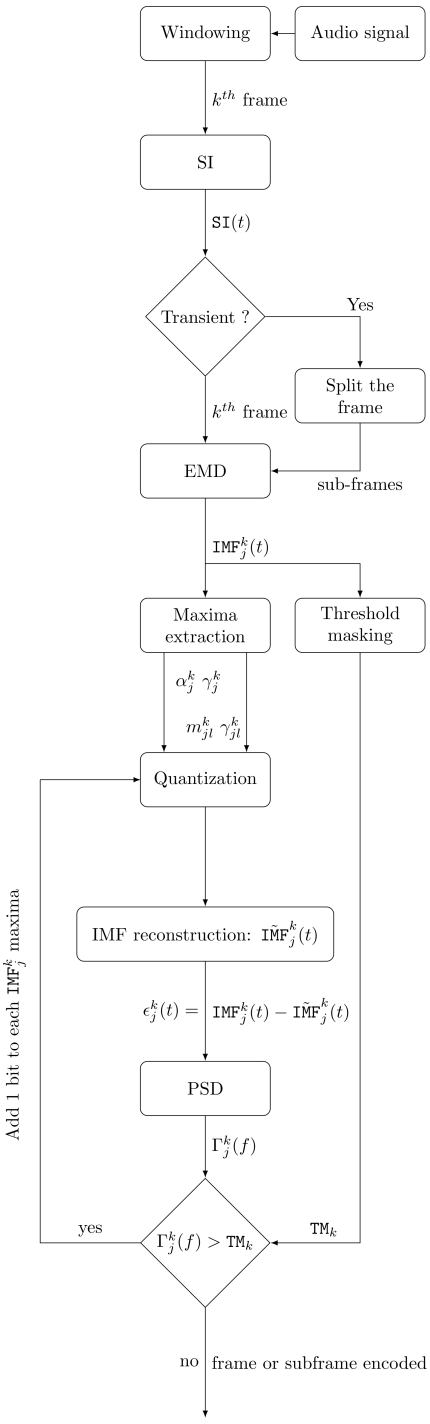

Figure 1: Synopsis of the EMC.

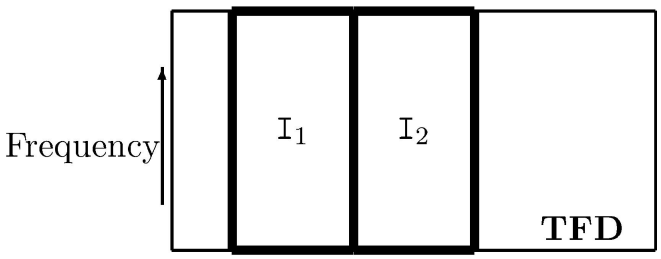

Figure 2: Domains $\mathtt{NI}_1$ and $\mathtt{NI}_2$ domains where $\mathtt{SI}$(t) is calculated.

stant t and extracted from $\rho(t, f)$:

$$\mathtt{I}_1(t;\tau,f) = \rho(t-L+\tau,f) \quad (2)$$

$$\mathtt{I}_2(t;\tau,f) = \rho(t+\tau,f) \quad (3)$$

where L is the time width of the sub-images and $\tau \in [0, L]$. The SI is obtained by computing the Bhattacharyya distance between the two sub-images as follows:

$$\mathtt{SI}(t) = -\log\Big(\int_{\tau=0}^{L}\int_{-\infty}^{+\infty}\sqrt{\mathtt{NI}_1(t;\tau,f)\mathtt{NI}_2(t;\tau,f)}df d\tau\Big) \quad (4)$$

where $\mathtt{NI}_k$ $(k = 1, 2)$ is the normalized version of the sub-image $\mathtt{I}_k$:

$$\mathtt{NI}_k(t;\tau,f) = \frac{\mid \mathtt{I}_k(t;\tau,f) \mid}{\int_{\tau=0}^{L}\int_{-\infty}^{+\infty} \mid \mathtt{I}_k(t;\tau,f) \mid df d\tau} \quad (5)$$

In this work the spectrogram is used as a TFD due to its very simple use and the absence of cross-terms. A peak in $\mathtt{SI}(t)$ shows rapid or abrupt change in the signal spectrum and thus indicates the presence of a transition zone. For the SI evaluation (Eq. 4), we work with sampled signals and for the numerical integration, Newton integration is used. Figure 3 shows an example of the variations of $\mathtt{SI}(t)$ for a "song" audio frame for different values of L in Eq. 5. This figure also shows that, for $L = 64$, the main transition is well evidenced by the SI. In practice, signals are split into overlapping frames of constant length. Figure 4 shows an example of a transient detection evidenced in an audio frame by the SI followed by the splitting (segmentation) of the frame into two overlapping sub-frames.

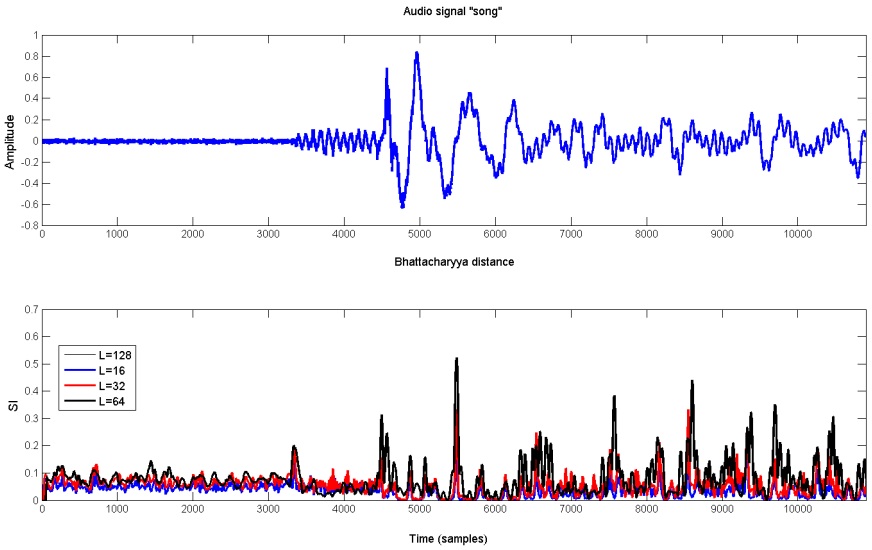

Figure 3: Stationarity index of an audio frame "song" [27].

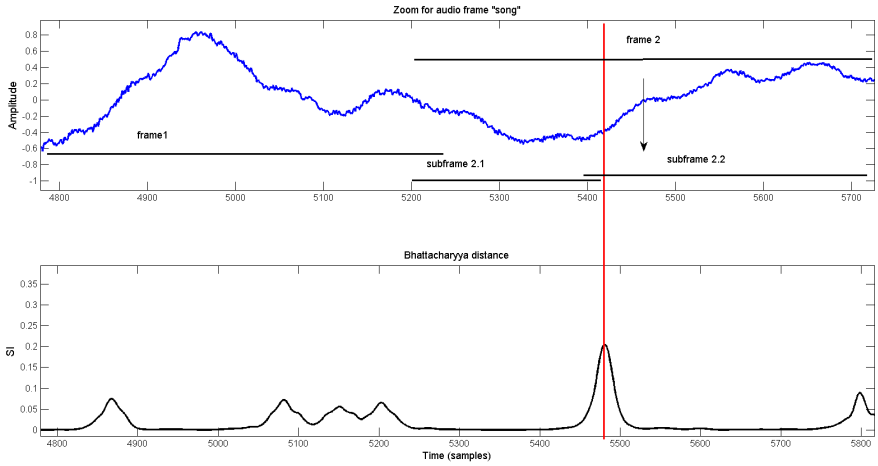

Figure 4: Example of segmentation for an audio frame "song" [27].

3.4. IMFs symmetry

One of the goals of the sifting process is to remove the asymmetry between the upper (maxima) and lower (minima) envelopes in order to transform the input signal into an AM-FM component. However, the EMD sifting is a numerical approach that may prevent the IMFs to be truly zero mean. This can be caused

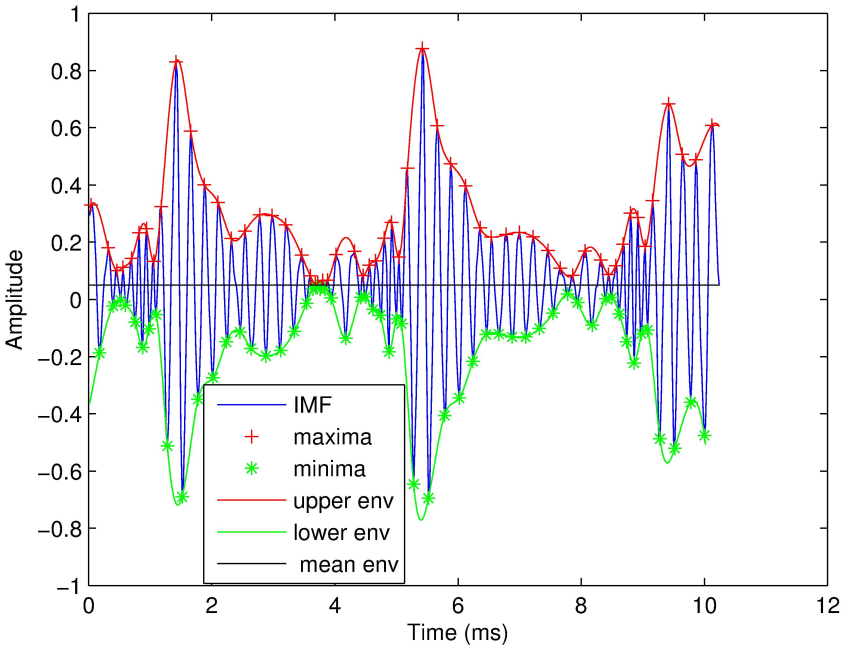

Figure 5: Example of IMF mean envelope offset.

by edge effects due to the construction of the envelopes through interpolation. The stopping criterion SD prevents from over decomposing the input signal but can also lead to the presence of offsets in IMFs. Thus, in general, the sifted IMFs are not truly symmetric with respect to the time axis (α_j^k=0) but they are symmetric about a shifted time axis, say $y = \alpha_j^k$. This fact is illustrated in figure 5 where the envelopes are symmetric with respect to line $y = 0.05$. Thus, provided the offset α_j^k is encoded, at the decoder the upper (resp. lower) envelope can be reconstructed and the lower (resp. upper) envelope determined by symmetry about the line $y = \alpha_j^k$. An example of offset values of five IMFs extracted from an audio frame signal are presented in Table 1. Note that the offset values depend on the amplitudes of the extracted IMFs. As expected, the IMFs are not all truly symmetric with respect to $y = 0$.

Table 1: Offset values of the IMFs extracted from an audio frame.

$\mathtt{IMF}_j$	1	2	3	4	5
$\boldsymbol{\alpha_j^k}$	0.05	0.02	0	0	0.006

3.5. Bit-allocation

A great attention has been paid to the problem of bit-allocation, where a given quota of bits is needed to efficiently allocate a certain number of bits to encode each audio frame [28],[29],[30]. A bit-allocation strategy consists in distributing, in a dynamic way, this fixed pool of bits over a number of signal components quantizers so that the audibility of the quantization process is minimized [29]. This results in an optimized audio quality for a given number of bits. The proposed bit-allocation strategy used to encode the maxima of the sifted IMFs and the residual, is subject to the following constraints:

1. The number of bits used to encode the maxima must be as small as possible.

2. The distortion between the true IMF and the reconstructed one must be as inaudible as possible.

To reduce the BR, a perceptual coding controlled by a psychoacoustic model is used to encode the amplitudes of the scaled maxima. The same psychoacoustic model as in MPEG-1 audio coder, involving the signal to Noise Mask Ratio (NMR) [4], is used. Initially, the number of allocated bits is fixed according to the BR coding. The number of bits allocated to each IMF is adjusted in order to ensure that $\Gamma_j^k(f) < \mathtt{TM}_k(f)$ [31]. Since each IMF contains less frequency oscillations (extrema) than each previously sifted ones, we start by quantizing the last IMF, which has the smallest number of extrema and therefore requires the fewer bits. Note that in general the first IMF (high frequency component) is not particularly smooth and its number of extrema is very large.

The way of distributing bits

Let N be the number of frames of constant duration. Each frame is broken down into C_k IMFs where k is the frame index. Let B_T be the total number of bits allocated for data compression. In practice, some frames require more or less bits than the average number of bits, $B_F = B_T/N$, for their encoding. Indeed, the number of IMFs is frame dependent and thus the number of allocated bits is adjusted to the number of extracted IMFs of each frame. We first start with an equal bit-allocation, where each frame is assigned the same number of bits, B_F. If there are surplus bits in a given frame (these bits are put into the reservoir bits), the amount of pre-allocated bits is updated for the next one. The coder can only borrow additional bits donated from past frames and not from future frames. Since low frequencies of each frame are embedded in the last IMF, less bits are required for encoding its associated maxima and their positions. Sampled maxima m_{jl}^k and time positions t_{jl}^k represent a certain amount of data that must be encoded so as to fit within the target BR. Let b_j^k be the number of bits allocated to the j^{th} IMF of the k^{th} frame. The number of bits assigned to each maximum (m_{jl}^k, t_{jl}^k) is equal to $(b_j^k - 8)/N_j^k$, because 8 bits are reserved for encoding the offset α_j^k. Since a direct optimization is infeasible, bit-allocation is done in an iterative way. A loop is intended to quantize the amplitudes of the scaled maxima, to reconstruct the IMF, and then to compare the PSD $\Gamma_j^k(f)$ to $\mathtt{TM}_k(f)$: if $\Gamma_j^k(f) < \mathtt{TM}_k(f)$, the quantization is updated with an increased number of bits, until the masking constraint is satisfied. The bit-allocation strategy is shown in figure 7 where, for each IMF, the process starts by allocating one bit to each maximum of the IMF ($i = 1$). This is followed by the inaudibility evaluation. At each iteration of the process for $\mathtt{IMF}_j^k$ the number of bits is increased by one ($i = i + 1$) for each maximum (m_{jl}^k, t_{jl}^k) and the bit-allocation loop is stopped when the reconstruction error for the IMF respects the inaudibility constraint. Finally, the classical uniform scalar quantization is used, followed by the Huffman coding.

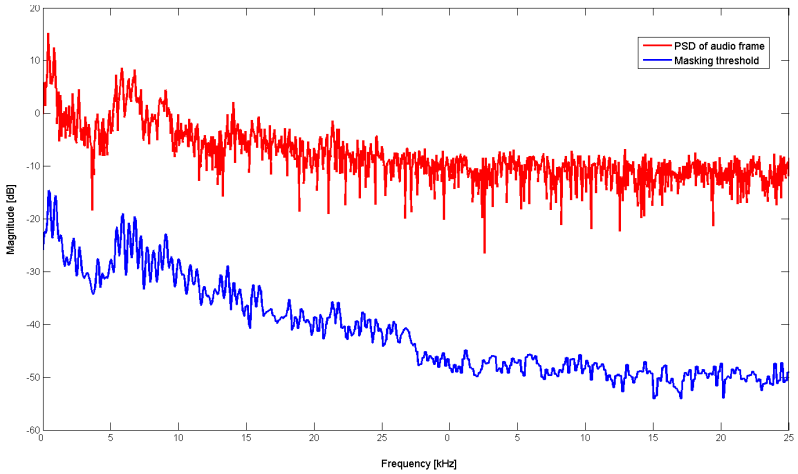

Figure 6: PSD of an audio frame and the associated masking threshold.

Note that the EMD method that decomposes a signal into a small number of narrow-band components (IMFs) which admit physically meaningful decomposition in terms of both instantaneous frequency and amplitude via the Hilbert transform or energy separation algorithm [32]. The term narrow-band refers both to the global signal properties, as defined in [33], and to the local signal requirements, as defined in [13]. This is consistent with the definition of a mono-component IMF. Applying the EMD to a frame (time-limited signals), the sifted IMFs might have infinite bandwidth. However, experiments carried out on a large class of signals, show that the power spectral densities of the obtained modes decrease fast outside a limited bandwidth [34]. In addition, it has been shown, based on extensive and controlled simulations, that the EMD exhibits dyadic filter bank properties when it is applied to the versatile class of fractional Gaussian noise processes (including white Gaussian noise) [35],[36].
As the MPEG-1 coder, there is one psychoacoustic evaluation per frame. Since the IMFs of a given frame have spectra with little overlap, the error spectrum for each frame can efficiently be compared directly to the threshold curve. The quality of real data experiments strengthens this analysis.

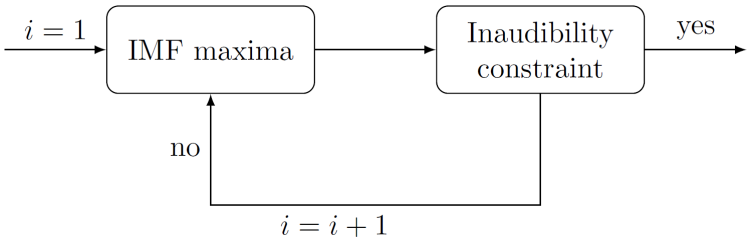

Figure 7: Bit-allocation strategy.

3.6. EMD decoder

The encoder gives a compact representation of the input audio signal that requires lower BR. For the EMC, this representation is given in terms of extrema of the IMFs. Coded extrema (m_{jl}^k and t_{jl}^k) of this compact representation are delivered to the decoder, which then recovers the original audio signal from the received compact representation. In conjunction with these extrema, the decoder uses side information such as the offset α_j^k and scale factor γ_j^k. To synthesize the audio signal $\tilde{x}(t)$, which is an approximation of the source $x(t)$, the decoder first reconstructs the upper envelope using γ_j^k and the encoded maxima m_{jl}^k and t_{jl}^k. Then the lower envelope of the j^{th} IMF is determined from the upper envelope by symmetry, with respect to the axis time, and using the decoded offset α_j^k (see subsection 3.4). Finally, the IMFs are recovered thanks to a spline interpolation between the extrema [19]. The audio frame is constructed by superposition of the estimated IMFs [13], and the decoded audio signal is obtained by frames concatenation. To hide the discontinuities at frame boundaries and for smooth transition, a cross-fading operation is included in the reconstruction step.

4. Results

4.1. Configuration parameters of the EMC

We begin by fixing the parameters of the EMC. The frame size is set to 512 with an overlapping of 64 samples. For the transient detection, the size L of sub-images, $\mathtt{NI}_k$, is set to 64. To best of our knowledge, there is no criterion

for optimum selection of L parameter [26]. This value is application dependent. Based on extensive simulations (large classes of audio signals) we found that, for frames of length 512, $L = 64$ is a good choice for change detection, as illustrated in figure 3 that shows the variations of the SI across the time for different values of L. To decide whether there is a non-stationarity a threshold ($\mathtt{Th}_{\mathtt{NS}}$), set to 10% of the highest peak of the SI, is used. Time index t_{jl}^k, scaling factor γ_j^k and offset value α_j^k are coded over 8 bits.

4.2. Measurement scores

In addition to listening tests, performance is analyzed using the BR, the NMR, the Subjective Difference Grade (SDG) [37], and the Objective Difference Grade (ODG) [38], which is a perceptual criterion. The SDG is the difference grade between listener's rating of the coded signal and the reference signal [39],[40]:

$$\text{SDG} = \text{Grade(coded)} - \text{Grade(reference)}$$

Values of the SDG range from -4 to 0 with the following interpretation, (-4) unsatisfactory (or) very annoying, (-3) poor (or) annoying, (-2) fair (or) slightly annoying, (-1) good (or) perceptible but not annoying, and (0) excellent (or) imperceptible. A group of seven listeners (randomly selected) have evaluated the analyzed audio coders. The average SDG was computed for each of the audio signals. The NMR is an objective measure of the perceptual quality of a compressed signal which measures the relative level of the quantization noise compared to the masking threshold [41]. Lower coding errors are indicated by larger negative values of the NMR. It has been shown that the NMR is a useful tool in the development and comparison of perceptual coding schemes. The NMR of a given sub-band is calculated for b-bit quantization as the difference in dB by,

$$\text{NMR}(b) = \text{SMR} - \text{SNR}(b)$$

where SMR is the signal-to-mask ratio calculated with the psychoacoustic model and SNR(b) is the signal-to-noise ratio, resulting from b-bits quantization, esti-

mated from a table lookup based on the number of bits allocated to the sub-band. Thus the NMR(b) value corresponds to the difference between the level of quantization noise (SMR) and the level (SNR(b)) where a distortion may just become audible in the given sub-band. The MPEG/audio standard provides tables that give estimates for the SNR, resulting from quantizing at a given number of quantization levels. For the psychoacoustic model, the same model as in MPEG-1 audio coder [4] is used. The ODG represents the expected human perceptual quality of the degraded signal. It is generated by a procedure designed to be comparable to the SDG judged by human ears. This score is calculated based on the difference between the quality rating of the reference signal and the test signal. This score takes its values from -4 to 0 where -4 stands for very significant difference and 0 stands for imperceptible difference between the reference and the test signal [42],[43],[44]. More precisely, values of 0,-1,-2,-3,-4 correspond to a subjective audio quality of "indistinguishable from original", "perceptible but not annoying", slightly annoying", annoying" and "very annoying" respectively.

4.3. Application to real signals

We present some experiments to assess the performance of the EMC method. The EMC is tested on six audio signals (gspi=Glockenspiel, harp=Harpsichord, quar=Quarter, song, Violin, trpt=Trumpet and violin) sampled at 44.1kHz. In particular, gspi, harp, quar and trpt recordings are taken from the SQAM database [47]. Listening tests required to ascertain the quality of the reconstructed audio relative to state-of-the art codecs and prove the relevance of the EMC are provided with the paper. Also, results are compared to those of MP3 (ISO/IEC 11172-3 MPEG Layer 3) [20],[45] and AAC (ISO/IEC 13818-7 Advanced Audio Coding) [46],[21], and to the wavelet-based coding [48]. Daubechies wavelets that are, widely used in solving a broad range of problems, are compactly supported orthonormal wavelets and provide accurate signal decomposition; hence they are good candidates for signals coding. In general, they achieve good results in audio coding compared to other wavelets [48]. In

this work, Daubechies wavelet Db8 is used as mother wavelet which yields orthogonal decomposition of the signal. For each frame a perceptual masking threshold is calculated. Spectrograms of the tested audio signals are depicted in figure 8. This figure shows that the audio signals exhibit varying and complex time-frequency structures. Based on the analyzed audio signals, the extracted residues of the decompositions ($\mathtt{r}_C(t)$) are of very small amplitudes (tend to zero) and thus have not been coded. Figure 9 illustrates the sifting of an audio signal frame. According to this decomposition and as expected [13]-[19] the number of maxima decreases from one IMF to IMF. Values of the NMR, the BR, the ODG and the SDG obtained at BR=64 kb/s with the four compression methods are summarized in Table 2.

A careful examination of these results shows that for a constant BR, the EMC outperforms the MP3 and the wavelet-based coding, and on the average, performs better than the AAC codec in terms of ODG, SDG and NMR. Compared to the wavelet-based coding and the MP3 codec, for all signals both the proposed coding and the AAC codec have the ODG values between -1 (not annoying) and 0 (not perceptible). These results show that, for the tested signals, the coding performance of the EMC does not show significant perceptual distortion. In terms of NMR, it can be observed from Table 2, that on the average, larger negative values are obtained with the EMC, leading to lower coding errors and less audible noise. Audio quality of the coding is also assessed in term of impairment using SDG scores. Compared to other three codecs, the SDG values of the EMC are near zero, showing the high quality of the decoded signals that are indistinguishable from the original signals. We observe that the EMC and the AAC perform better than the MP3 and the wavelet-based coding methods in terms of SDG. We report in Table 3 a result showing the interest of encoding the offset values. This result shows that the ODG is improved when the offset is coded. Overall, when compared to the MP3, the AAC and the wavelet-based coding, there is a preference toward the EMC for the six tested audio signals. The good behavior of the EMC compared to these coding techniques lies in the fact that, although they can be compressed at low BR expense, the IMFs are

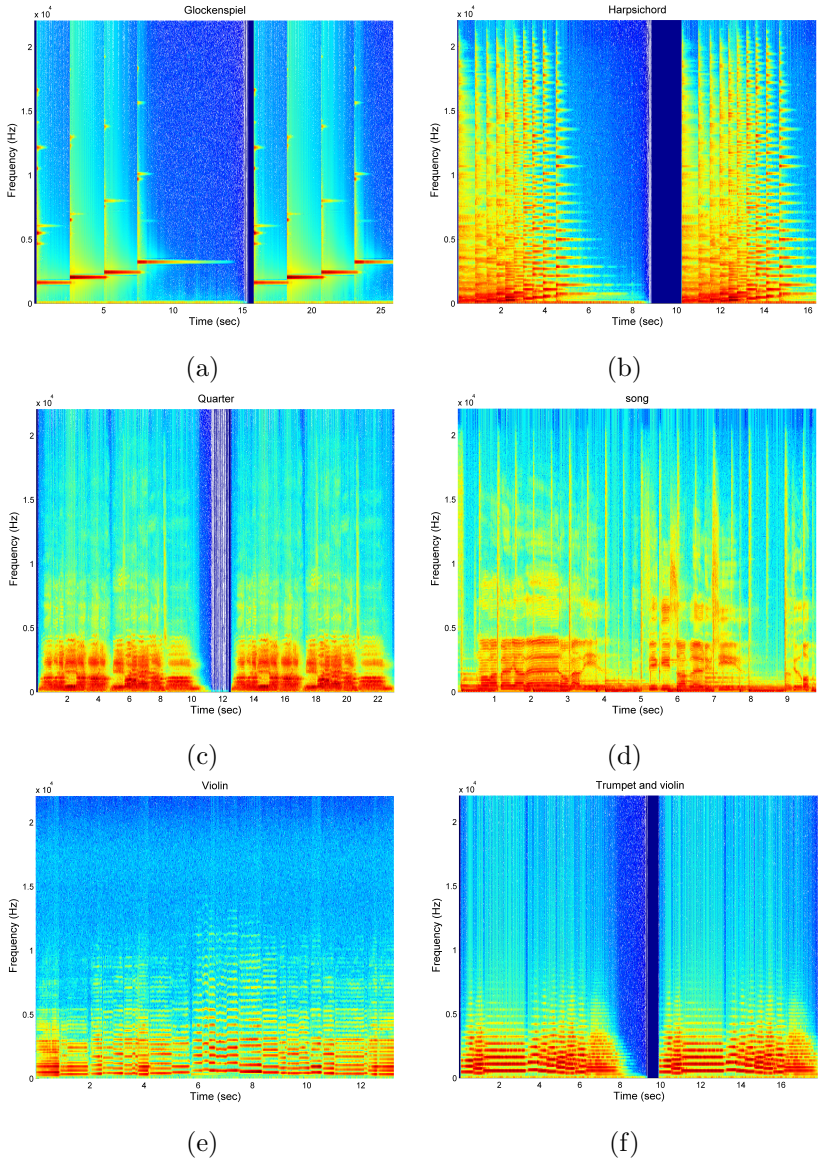

Figure 8: Original audio signals (Glockenspiel, Harpsichord, Quarter, song, Violin, Trumpet and violin).

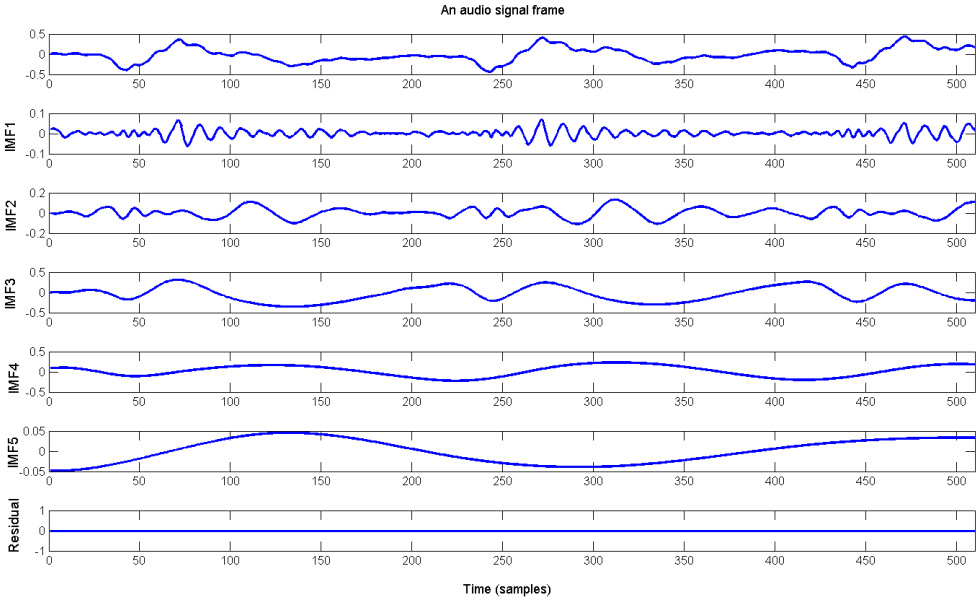

Figure 9: Example of decomposition of an audio signal frame.

able to catch easily some non-stationary behavior as well as several harmonics of the signal that require more complexity with approaches focused on harmonic retrieval or other fixed bases decomposition. For instance, in figure 9, we see that the first IMF catches several stationary periodicities while the 4^{th} IMF catches a single harmonic with slowly varying period (the first period is about 175 samples long and the second 190). This frequency variation is even more clearer in 5^{th} IMF.

The high compression achieved by the EMC is mainly attributed to data-driven nature of the EMD, to the use of a psychoacoustic model and to the exploitation the symmetry property of the IMF, which enable good audio quality at low BR. Although the processed test signals have different and varying time-frequency structures (Fig. 8), the sifting is performed without any prior choice of basis functions and the number of extracted IMFs is also adaptively driven. Based on a psychoacoustic model, the aim of this new coding is to keep the listening

Table 2: Compression results for audio signals (gspi, harp, quar,song, trpt and violin) by the proposed EMC approach, AAC, MP3 and wavelet compression.

	Signal	gspi	harp	quar	song	trpt	violin
EMC	BR [kb/s]	64	64	64	64	64	64
	NMR	-5.37	-5.65	-5.47	-5.13	-5.32	-5.04
	ODG	-0.82	-0.73	-0.74	-0.79	-0.84	-0.83
	SDG	-0.71	-0.91	-0.86	-0.75	-0.87	-0.78
AAC	BR [kb/s]	64	64	64	64	64	64
	NMR	-3.43	-6.46	-4.78	-4.23	-6.15	-4.59
	ODG	-0.85	-0.73	-0.75	-0.89	-0.88	-0.86
	SDG	-0.77	-0.85	-0.88	-0.82	-0.87	-0.81
MP3	BR [kb/s]	64	64	64	64	64	64
	NMR	1.42	1.21	1.27	1.23	2.68	1.86
	ODG	-1.12	-1.87	-1.91	-1.09	-1.27	-1.34
	SDG	-1.08	-1.22	-1.52	-0.96	-1.08	-1.21
Wavelets	BR [kb/s]	65	67	64	65	66	64
	NMR	-2.30	-3.67	1.64	-3.40	-1.35	-2.52
	ODG	-0.86	-1.27	-1.74	-0.98	-0.97	-1.08
	SDG	-0.93	-1.13	-1.40	-0.93	-0.97	-1.03

Table 3: EMC with and without offsets coding.

EMC	ODG
With offset coding	-0.82
Without offset coding	-0.97

quality of the signal at a consistent level. Performance of the EMC depends on the quality of the sifting which in turns depends on the way the interpolation of the envelopes is performed. Thus, utilizing an inappropriate interpolating

function can limit the performances of the EMD-based coding scheme. Note that the spline interpolation of the extrema can introduce errors in the reconstructed signal. This is particularly true for the first IMF which holds the most non-smooth part of the signal [18]. The EMC is based on the EMD which is essentially defined by the sifting procedure and involves four main steps: Extrema detection, Interpolation, Sifting and Reconstruction. Extrema detection and interpolation requires, on average, about 50% of the computational time. The computational complexity of the EMD depends on the data (signal) to be sifted, the chosen interpolation scheme, and the stopping criterion. For example, the distribution of local maxima becomes more dense when the number of samples of the analyzed signal increases. Since the EMD is lacking an analytical definition, algorithmic complexity of both the EMD and the EMC are difficult to evaluate.

The present version of the EMC is limited to BR greater than 64 kb/s since the bit-allocation is based on the comparison to the perceptual masking threshold (inaudibility constraint). For BR $\geq$ 64 kb/s, the encoder does not run out of bits. There are enough bits to encode all the maxima. Based on the EMD, the EMC is data driven approach and thus the BR allocation is signal dependent. It is important de keep in mind that before coding, the number of IMFs per frame and the associated number of maxima are unknown. As the EMD, the bit-allocation is also, in big part, a data driven process. For significantly lower BR ($\leq$ 64 kb/s) the iterations are stopped and the EMC does not give the expected coding results. Since the sifting depends on the `SD` value, it is expected that for some frames few IMFs of no zero-mean would be extracted. An example of such IMFs is illustrated in Table 1. However, the BR is improved for the IMFs of zero-mean because the offset is null and thus it is not coded (the residual is inaudible). To support very low BRs, an improvement of the EMC is necessary while keeping a good listening quality of audio signals.

5. Conclusion

This paper presents a new approach for audio coding based on the EMD. Experimental results demonstrate the effectiveness of such a framework compared to classical approaches. We have compared the EMC with the MP3, the AAC and the wavelet-based coding. Overall, based on the analyzed mono audio signals, the EMC performs better than these methods for the tested signals: ODG values obtained for the EMC method are better than with the three codecs without significant perceptual distortion, and large negative values of the NMR obtained with the EMC lead to lower coding errors and less audible noise. SDG scores near to zero obtained for the EMC indicate the high quality of the decoded signals, that are indistinguishable from the original ones. The coding results of the EMC are not dependent on predetermined basis and/or sub-band filtering processes. Furthermore, it does not require any user parameters, except the window size and the threshold for the stationarity index calculation. In future work, we plan to design a strategy to optimize the selection of these parameters. Simulations presented here are restricted to BR greater than 64kb/s and in future work we also plan to extend the EMC for operation at lower BR like 48-32 kbits/s, while keeping a good listening quality of audio signals. Practical experiments carried out here on different kinds of audio sources should be extended by considering larger classes of audio signals as well as varied experimental conditions such as different sampling rates or frame size for further performance improvement. Since it is based on the EMD, the EMC shares the same limits: it is only defined by an algorithm (sifting) and lacks clear mathematical framework (with the notable exception of [49]) and further work in this direction would be welcome. Future works should also compare the performance of the EMC to other coding methods involving MDCT basis functions. Since the EMC is a mono coding approach, investigations are required for its extension to stereo audio signals.

[1] N. Jayant, "Signal compression," *Int. J. High Speed Electron. Syst.,* vol. 8, no. 1, pp. 1-12, 1997.

[2] R.N.J. Veldhuis, M. Breeuwer, R.G. Van Der Waal, "Subband coding of digital audio signals," *Phillips J. Res.,* vol. 44, pp. 329-343, 1989.

[3] J.D. Johnston, "Transform coding of audio signals using perceptual criteria," *IEEE J. Select Areas Commun.,* vol. 6, pp. 314-323, 1988.

[4] P. Noll, "MPEG digital audio coding," *IEEE Sig. Proc. Mag.,* vol. 14, no. 5, pp. 59-81, 1997

[5] E. Ravelli, G. Richard and L. Daudet, "Union of MDCT bases for audio coding," *IEEE Trans. Speech Audio Proc.,* vol. 16, no. 8, pp. 1361-1372, 2008.

[6] T.K. Truong, P.D. Chen and T.C. Cheng, "Fast algorithm for computing the forward and inverse MDCT in MPEG audio coding," *Sig. Proc.,* vol. 86, pp. 1055-1060, 2006.

[7] A. Petrovsky, E. Azarov and A. Petrovsky, "Hybrid signal decomposition based on instantaneous harmonic parameters and perceptually motivated wavelet packets for scalable audio coding," *Sig. Proc.,* vol. 91, pp. 1489-1504, 2011.

[8] N. Ruiz, P.V. Candeas and F.L. Ferreras, "Wavelet-based approach for transient modeling with application to parametric audio coding," *Digital Sig. Proc.,* vol. 20, pp. 123-132, 2010.

[9] N. Ruiz, M. Rosa, F. Lopez and P. Jarabo, "Adaptive wavelet-packet analysis for audio coding purposes," *Sig. Proc.,* vol. 83, no. 5, pp. 919-929, 2003.

[10] G. Stoll, S. Nielsen and L. Van de Kerkhof, "Generic architecture of the ISO/MPEG audio layer I and II-Compatible developments to improve quality and addition of new features," *Conv. Aud. Eng. Soc.,* 1993

[11] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.,* vol. 42, no. 10, pp. 780-792, 1994.

[12] D. Sinha and A. Tewfik, "Low bit rate transparent audio compression using adapted Wavelets," *IEEE Trans. ASSP,* vol. 41, no. 12, pp. 3463-3479, 1993.

[13] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung and H.H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Royal Society,* vol. 454, no. 1971, pp. 903-995, 1998.

[14] K. Khaldi and A.O. Boudraa, "Audio watermarking via EMD", *IEEE Trans. Audio, Speech and Language Proc.,* vol. 21, no. 3, pp. 675-682, 2013.

[15] K. Khaldi, A.O. Boudraa and A. Komaty, "Speech enhancement using empirical mode decomposition and Teager-Kaiser energy operator," *Journal of Acoustical Society of America,* vol. 135, no. 1, pp. 451-459, 2014.

[16] K. Khaldi, A.O. Boudraa and M. Turki, "Voiced/unvoiced speech classification-based adaptive filtering of decomposed empirical modes for speech enhancement," *IET Sig. Proc.,* vol. 10, Iss. 1, pp. 69-80, 2016.

[17] K. Khaldi and A.O. Boudraa, "On signals compression by EMD," *IEE Electronics Lett.*, vol. 48, issue 21, pp. 1329-1331, 2012.

[18] A. Linderhed, "2D empirical mode decompositions in the spirit of image compression," *in Wavelet and Independent Component Analysis Applications IX , SPIE Proceedings,*, vol. 4738, pp. 1-8, 2002.

[19] K. Khaldi, A.O. Boudraa, M. Turki, Th. Chonavel and I. Samaali, "Audio encoding based on the empirical mode decomposition," *EUSIPCO,* Glasgow, Scotland, pp. 924-928, 2009.

[20] ISO/IEC 11172-3 (Information Technology Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5 Mbit/s)−Part 3: Audio, International Organization for Standardization, 1993.

[21] http://www.apple.com/itunes/

[22] J.P. Bello, L. Daudet, S. Abdallah, Ch. Duxbury, M. Davies, and M.B. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Trans. Audio, Speech and Language Proc.,* vol. 13, no. 5, pp. 1035-1047, 2005.

[23] M. Basseville and I. V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[24] N. Wachowski and M.R. Azimi-Sadjadi, "Detection and Classification of Nonstationary Transient Signals Using Sparse Approximations and Bayesian Networks," *IEEE/ACM Trans. Audio, Speech, and Language Proc.,* vol. 22, no. 2, pp. 1750-1764, 2014.

[25] V. Bruni, S. Marconi, D. Vitulano, "Time-scale Atoms Chains for Transients Detection in Audio Signals", *IEEE Trans. Audio, Speech and Language Proc.,* vol. 18, no. 3, pp. 420-433, 2010.

[26] H. Laurent and C. Doncarli, "Stationarity index for abrupt changes detection in the time frequency plane," *IEEE Sig. Proc. Lett.,* vol. 5, no. 2, pp. 43-45, 1998.

[27] *http://michel-jonasz.freedownloadmp3.net/chanson-francaise*

[28] Y. Shoham and A. Ghersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. ASSP,* vol.36 , no. 9, pp. 1445-1453, 1988.

[29] S. Voran, "Perception-based bit-allocation algorithms for audio coding," *Proc. IEEE Workshop on ASSP,* pp. 1-4, 1997.

[30] V.K. Goyal, "Theoretical foundations of transform coding," *IEEE Sig. Proc. Mag.,* vol. 18, no. 5, pp. 9-21, 2001.

[31] K. Khaldi, A.O. Boudraa, M. Turki and Th. Chonavel, "Codage audio perceptuel bas débit par décomposition en modes empiriques (EMD)," *GRETSI,* Dijon, France, 2009.

[32] A.O. Boudraa and F. Salzenstein, "TeagerKaiser energy methods for signal and image analysis: A review," *Digital Sig. Proc.,* vol. 78, pp. 338-375, 2018.

[33] M. Schwartz, W.R. Bennet and S. Stein. Communications systems and techniques. McGraw-Hill, 1966.

[34] Hilbert-Huang Transform and its Applications, N.E. Huang and S.S.P. Shen editors, 324 pages, World Scientific Publishing, 2011.

[35] P. Fladrin, G. Rilling and P. Gonçalves, "Empirical mode decomposition as a filter bank," *IEEE Sig. Proc. Lett.,* vol. 11, no. 2, pp. 112-114, 2004.

[36] P. Flandrin and P. Gonçalves, "Empirical mode decompositions as data-driven wavelet-like expansions," *Int. J. Wavelets, Multiresolution and Information Proc.,* vol. 2, no. 4, pp. 477-496, 2004.

[37] R. Huber and B. Kollmeier, "PEMO-QA New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio, Speech and Language Proc.,* vol. 14, no 6, pp. 1902-1911, 2006.

[38] Method for Objective Measurements of Perceived Audio Quality, *ITU Recommendation, ITU-R BS.1387-1*, 2001.

[39] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE,* vol. 88, no. 4, pp. 451-515, 2000.

[40] "ITU-R Rec. BS.562: Subjective assessment of sound quality," Int. Telecomm. Union, Geneva, Switzerland, 1990.

[41] K. Brandenburg and T. Sporer, "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria," Proc. AES 11th Int. Conf. Test and Measurement, pp. 169-179, 1992.

[42] Method for Objective Measurements of Perceived Audio Quality, Draft ITU-T Recommendation BS.1387, Jul. 2001.

[43] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality (Tech. Rep.)". Montreal, Canada: McGill University, Department of Electrical and Computer Engineering, 2003.

[44] C.H. Yang and H.M. Hang, "Cascaded Trellis-Based Rate-Distortion Control Algorithm for MPEG-4 Advanced Audio Coding," *IEEE Trans. Audio Speech and Language Proc.,* vol. 14, no. 3, pp. 998-1007, 2006.

[45] http://www.iis.fraunhofer.de/bf/amm/

[46] ISO/IEC 13818-7 (MPEG-2 Advanced Audio Coding, AAC), International Organization for Standardization, 1997.

[47] Sound Quality Assessment Material recording for subjective tests, *Technical Centre of the European Broadcasting Union,* 1988.

[48] M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice Hall Signal Processing Series (1995).

[49] E.H.S. Diop, R. Alexandre and A.O. Boudraa, "Analysis of Intrinsic Mode Functions: A PDE approach," *IEEE Sig. Proc. Lett.*, vol. 17, no. 4, pp. 398-401, 2010.