HHT-based audio coding

Kais Khaldi, Abdel-Ouahab Boudraa, Bruno Torrésani, Thierry Chonavel

▶ **To cite this version:**

Kais Khaldi, Abdel-Ouahab Boudraa, Bruno Torrésani, Thierry Chonavel. HHT-based audio coding. Signal, Image and Video Processing, 2013, 7, pp.1-9. 10.1007/s11760-013-0433-6 . hal-00818033

HAL Id: hal-00818033
https://hal.science/hal-00818033v1

Submitted on 25 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HHT-based audio coding*

Kais Khaldi[1,2], Abdel-Ouahab Boudraa[1], Bruno Torresani[3] and Thierry Chonavel[4]

Abstract

In this paper a new audio coding scheme combining the Hilbert transform and the Empirical Mode Decomposition (EMD) is introduced. Based on the EMD, the coding is fully data-driven approach. Audio signal is first decomposed adaptively, by EMD, into intrinsic oscillatory components called Intrinsic Mode Functions (IMFs). The key idea of this work is to code both instantaneous amplitude (IA) and instantaneous frequency (IF), of the extracted IMFs, calculated using Hilbert transform. Since IA (resp. IF) is strongly correlated, it is encoded via a linear prediction technique. The decoder recovers the original signal by superposition of the demodulated IMFs. The proposed approach is applied to audio signals, and the results are compared to those obtained by AAC (Advanced Audio Coding) and MP3 codecs, and wavelets based compression. Coding performances are evaluated using the bit rate, Objective Difference Grade (ODG) and Noise to Mask Ratio (NMR) measures. Based on the analyzed audio signals, overall, our coding scheme performs better than wavelet compression, AAC and MP3 codecs. Results also show that this new scheme has good coding performances without significant perceptual distortion, resulting in an ODG in range [-1,0] and large negative NMR values.

Index Terms

Audio coding, Empirical mode decomposition, Intrinsic mode function, Hilbert transform, Hilbert-Huang transform, Linear prediction.

*Preliminary results of this work has been presented at IEEE ISCCSP conference, Limassol, Cyprus, 2010.

[1]IRENav, Ecole Navale, BCRM Brest, CC 600, 29240 Brest Cedex 9, France.
[2]U2S, ENIT, BP 37, Le Belvédère 1002, Tunis, Tunisia.
[3]LATP, CMI, Université de Provence, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France.
[4]LabSTICC UMR, Télécom Bretagne, BP 832, 29285 Brest Cedex, France.

I. Introduction

Signal coding is a well known problem in signal processing and particularly, in the case of audio signals where a number of methods have been proposed for reducing the Bit Rate (BR) requirements [1]-[2]. For more BR reduction with high fidelity, different approaches have been proposed [3]-[8]. Several approaches, that involve pre-determined basis functions (cosine, Daubechies,...), yield better results in terms of BR. Unfortunately, using fixed basis functions prevents the decomposition from being parsimonious for any kind of signals. As a matter of fact, even if a basis is well suited for a class of signals, in the sense that it yields compact descriptions with only a few significant terms, there are other signals for which the basis under consideration performs poorly. Thus, there is a need for data driven coding approach. Recently, the Empirical Mode Decomposition (EMD) has been introduced for analyzing non-stationary data derived from linear or non-linear systems in totally adaptive way [9]. This new expansion decomposes adaptively any signal into intrinsic oscillatory components called Intrinsic Mode Functions (IMFs). These extracted modes are zero-mean with symmetric envelopes AM-FM components. Basis functions of EMD are derived from the signal itself and hence, the decomposition is adaptive in contrast to traditional methods where the basis functions are fixed. A salient property of the IMF is that it can be fully described by its extrema. We have recently shown that the EMD can be used for audio signals compression by coding extrema of the IMFs and their positions [10]. Compared to our previous approach [11], the key idea of the present work is to code both instantaneous amplitude (IA) and instantaneous frequency (IF) of the extracted IMF. To further reduce the BR, we exploit the fact that both values of IA and IF of the IMFs are strongly correlated. Thus, each IA (IF) value is closely approximated as a linear combination of its past values. For each mode, the IF component is fitted with an Auto Regressive (AR) model, the order

of which is selected from the values of the partial autocorrelation coefficient of the extracted mode. Compared to instantaneous phase encoding of the extrema that requires higher BR (96kb/s) [11], the proposed predictive coding reduces the BR by exploiting the correlation of IF values. For the IA component, the order of the model is selected from a perceptual constraint [11]. More precisely, the psycho-acoustic model rules the selection of this order and keeps the listening quality of audio signal at a consistent level [12].

Our contribution can be viewed as a proof-of-concept of EMD based encoding for audio signals. The proposed scheme is applied to audio signals, and the results are compared to MP3 (MPEG-1/2 Audio Layer 3) codec and wavelets approach. Compared to our previous work [11] we include the AAC (Advanced Audio Coding) codec. Coding performances are evaluated using BR, objective difference grade and noise to mask ratio. The paper is organized as follows: section II introduces the Hilbert and Huang transforms. Section III presents the proposed coding approach and section IV describes the experimental results.

II. Hilbert-Huang Transform (HHT)

The EMD decomposes any signal $x(t)$ into a series of IMFs through an iterative process called *sifting*; each one with a distinct time scale [9]. The decomposition is based on the local time scale of $x(t)$, and yields adaptive basis functions. The EMD can be seen as a type of wavelet decomposition whose subbands are built up as needful to separate the different components of $x(t)$. Each mode replaces the signals detail, at a certain scale or frequency band [13]. The EMD picks out iteratively the highest frequency oscillation that remains in $x(t)$. The IMFs are extracted subject to two requirements:

1. First, the number of extrema and the number of zeros crossings may differ by no more than one.

2. Second, the average value of the envelope defined by the local maxima, and the envelope defined by the local minima, is zero.

Locally, each IMF contains lower frequency oscillations than the previously extracted one [9]. To be successfully decomposed into IMFs, the signal $x(t)$ must have at least two extrema (one minimum and one maximum). An IMF is extracted using the sifting process as follows:

`Step 1`: **Fix** the threshold ϵ and set $j \leftarrow 1$ (j^{th} `IMF`)

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

`Step 3`: **Extract** the j^{th} `IMF` :

`(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 (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** : $\mathtt{h}_{j,i}(t) := \mathtt{h}_{j,i-1}(t) - \mu_{j,i-1}(t)$

`(f)` : **Calculate** the stopping criterion :

$$\mathtt{SD}(i) := \sum_{t=1}^{T} \frac{|\mathtt{h}_{j,i-1}(t) - \mathtt{h}_{j,i}(t)|^2}{(\mathtt{h}_{j,i-1}(t))^2}$$

`(g)` : $i := i + 1$

`(h)` : **Repeat** `Steps (b)-(g)` until $\mathtt{SD}(i) < \epsilon$ and then $\mathtt{IMF}_j(t) \leftarrow \mathtt{h}_{j,i}(t)$($j^{th}$ `IMF`)

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

`Step 5`: **Repeat** `Step 3` with $j := j + 1$ until number of extrema of $\mathtt{r}_j(t)$ is ≤ 2.

T is the time duration of $x(t)$. The sifting is repeated until the component $h_{j,i}(t)$ satisfies the aforementioned requirements (1) and (2). The signal $x(t)$ is finally written as the sum

of mode time series:

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

where $\mathtt{IMF}_j(t)$ is the IMF of order j, $\mathtt{r}_C(t)$ is the residual and C is the number of IMFs determined automatically using the stopping criterion (Standard Deviation) $\mathtt{SD} < \epsilon$. Usually, ϵ is set between 0.2 and 0.3 [9]. Figure 1 shows an example of decomposition of an audio frame by EMD. One can remark that the first IMF corresponds to fast oscillations, whereas the sixth IMF corresponds to slow ones (Fig. 1).

Using Hilbert transform (HT), IA $a(t)$ and IF $f(t)$ of each IMF are calculated as follows:

$$\begin{aligned} a(t) &= \sqrt{[\mathtt{IMF}(t)]^2 + \mathcal{H}^2[\mathtt{IMF}(t)]} \\ f(t) &= \frac{1}{2\pi}\frac{d\theta(t)}{dt} \\ \theta(t) &= \tan^{-1}\left(\frac{\mathcal{H}[\mathtt{IMF}(t)]}{\mathtt{IMF}(t)}\right) \\ \mathcal{H}[\mathtt{IMF}(t)] &= \frac{1}{\pi}\mathtt{PV}\int_{-\infty}^{+\infty}\frac{\mathtt{IMF}(\tau)}{t-\tau}d\tau \end{aligned}$$

where $\mathcal{H}[x(t)]$ is the HT of $x(t)$ and $\mathtt{PV}$ is the Cauchy principal value of the integral. Combination of EMD and HT is designated as Hilbert-Huang Transform (HHT). The idea of this work is to encode both IA and IF functions of the audio signal [14]-[15] using linear prediction model.

III. Coding principle

The proposed encoding is done frame by frame. Thus, the first step consists in splitting the audio stream into frames of constant length. Enframing the audio signal allows to treat each frame as a relatively stationary sound. To guarantee the stationarity of each segmented frame, a detector is used to measure the invariance of the statistical properties of the frame over the time. If a transient is detected, the frame is divided into two sub-frames [16]. In this work, detection of transient sequence is based on the Local Entropic Criterion (LEC)

which is a non parametric rupture detector. This criterion is calculated over sliding window. The LEC of signal $x(t)$ is given by [16]:

$$\mathrm{LEC}_x(t) = \frac{\mathrm{E}_{xc}(t) - (\mathrm{E}_{xl}(t) + \mathrm{E}_{xr}(t))}{|\mathrm{E}_{xc}(t)|} \tag{2}$$

where $\mathrm{E}_{xc}(t)$, $\mathrm{E}_{xl}(t)$ and $\mathrm{E}_{xr}(t)$ denote the entropies of the principal window and of the left and right sub-windows respectively and are given by

$$\begin{aligned} \mathrm{E}_{xc}(t) &= \mathrm{E}_{x[t-\frac{N}{2},t+\frac{N}{2}-1]} \\ \mathrm{E}_{xl}(t) &= \mathrm{E}_{x[t-\frac{N}{2},t-1]} \\ \mathrm{E}_{xr}(t) &= \mathrm{E}_{x[t,t+\frac{N}{2}-1]} \end{aligned} \tag{3}$$

where $\mathrm{E}_{x[0,N-1]}$ is the Shannon entropy of $x(t)$ calculated over the interval $[0, N-1]$ and defined by:

$$\mathrm{E}_{x[0,N-1]} = -\sum_{k=0}^{N-1} |X(k)|^2 \log |X(k)|^2 \tag{4}$$

with $X(k)$ the normalized discrete Fourier transform of $x(t)$. Thus, the LEC takes its values in the range $[-1, 1]$. A transient in the signal is characterized by a LEC greater than 0. An example of LEC variations across an audio frame is shown in figure 2, with N set to 64 [16]. Figure 3 shows an example of segmentation into two sub-frames of an audio frame of 1500 samples (zoom of the audio frame signal of figure 2) using the LEC.

A. IF and IA codings

It was found that for a large class of analyzed audio signals, both IA and IF values of IMFs are strongly correlated. So, the AR model is used to efficiently exploit this correlation. Thus, IA and IF components of each IMF are modeled as follows:

$$a(t) = \sum_{k=1}^{p} c_a(k) a(t-k) + \varepsilon_1(t)$$

$$f(t) = \sum_{i=1}^{q} c_f(i) f(t-i) + \varepsilon_2(t) \tag{5}$$

where $[c_a(1), c_a(2), \ldots, c_a(p)]$ and $[c_f(1), c_f(2), \ldots, c_f(q)]$ are the coefficients of the AR model for $a(t)$ and $f(t)$ respectively. $\varepsilon_1(t)$ and $\varepsilon_2(t)$ are two zero mean white noise processes. The coefficients are calculated by minimizing the mean square error criterion. In the proposed approach, we code the coefficients of the AR model and the noise variance. For each IMF, the model order of the associated IF (IA) is determined. For IA component, AR order selection is controlled by the psychoacoustic model [11]. The order is adjusted so that power spectral density (PSD) of the reconstruction error of the IMF does not exceed its masking curve. The interest of using the psychoacoustic model is also to improve the compression gain, while preserving the listening quality of the decoded signal. Since each IMF contains lower frequency oscillations than each previously extracted ones, the order of IF component varies from one IMF to next one. The selection of the order of the AR model is based on the estimation of the partial autocorrelation coefficients that fits the variations of the corresponding IF. An advantage of the proposed IF coding strategy is that it can support variable BR coding. Figure 4 shows the partial autocorrelation coefficient of IMFs of frame audio signal (Fig. 1). The value from which the partial autocorrelation curve is constant, is identified as the order for IF modeling (Fig. 4). The order of AR model of each IF is presented in Table I.

B. Bit allocation

Bit allocation is done frame-by-frame. Since the amount of a bit allocation is signal dependent, we start with an equal bit allocation where each frame is assigned the same number of bits. However, in practice, some frames need more or less bits than the average number of bits for their encoding. The input signal is segmented into M overlapping frames

(time windows) of constant length, L, and each one is splitted into C_k IMFs where k is the frame index. Let $\mathtt{B}_s^r$ be the BR of the coding and $\mathtt{F}_s$ the sampling frequency of the input audio signal. The number of bits assigned to each frame is given by $\mathtt{B}_F = \mathtt{B}_s^r \times L/\mathtt{F}_s$ and the number of bits per IMF is given by $\mathtt{B}_k = \mathtt{B}_F/C_k$. The total number of bits allocated for the input signal is then $\mathtt{B}_T = M \times \mathtt{B}_F$. Let p_k^l and q_k^l be the selected orders for IA and IF AR models of mode $\mathtt{IMF}_l$ of the k^{th} frame, where $l \in \{1, \ldots, C_k\}$. For each frame, we first start by quantizing the order values (p_k^l, q_k^l) each one is coded with m_o bits. The amount of bits assigned to each coefficient c_a or c_f, and variances ε_1 or ε_2, of $\mathtt{IMF}_l$, is given by

$$b_k^l = \frac{(B_k - 2m_o C_k)}{\sum_{k=1}^{C_k} (p_k^l + q_k^l + 2)} \tag{6}$$

where $(p_k^l + q_k^l + 2)$ is the total number of parameters of the model (Eq. (5)). Since the number of IMFs is frame dependent, the number of bits allocated is adjusted with the number of extracted modes of each frame (Eq. (6)). If there are surplus bits in a given frame, the amount of pre-allocated bits is updated for the next one. More precisely, if the surplus of the k^{th} frame, denoted by $\mathtt{S}_k$, is different from zero the total number of bits assigned to $(k+1)$ is update to $(\mathtt{B}_F + \mathtt{S}_k)$. The coder can only barrow bits donated from pas frames and not from future frames. All surplus bits constitute a "bit reservoir". The proposed coding operates on $\mathtt{B}_s^r \geq 64$ kb/s and using surplus bits, the coder will not run out of bits. There are enough bits to encode all coefficients.

C. Coding improvement

Coding of AR model coefficients and noise variance can be improved using lossless compression such as Huffman or Lempel-Ziv coding techniques to store data. These techniques account for probability of occurrence of encoded data to reduce the number of bits allocated

to store data. Although Lempel-Ziv is not optimum, the decoder does not require any coding dictionary [17].

D. Decoding process

The estimated $\hat{\texttt{IMF}}(t)$ of $\texttt{IMF}(t)$, for IF coding approach, is calculated as follows:

$$\hat{\texttt{IMF}}(t) = |\hat{a}(t)| \cos(\int_0^t 2\pi \hat{f}(\tau) d\tau) \tag{7}$$

where $\hat{a}(t)$ and $\hat{f}(t)$ are the estimates of $a(t)$ and $f(t)$ respectively using linear prediction approach. The audio frame is constructed by superposition of the estimated IMFs, and the decoded audio signal is obtained by frames concatenation.

E. **Phase initialization**

Since we are working with windowed signals (Hamming window) where the window amplitude decays progressively to zero at both ends, phase initialization does not matter very much. Indeed, phase progressively adapts to signal when moving from the beginning to the center part of the frame, where signal energy becomes non negligible. To check this, we have added a simulation where phase is initialized randomly. Denoting φ the selected initial phase and $x_{(\varphi)}$ the corresponding reconstructed signal, the relative reconstruction error, defined as

$$E_\varphi = \frac{\parallel x - x_{(\varphi)} \parallel^2}{\parallel x \parallel \times \parallel x_{(\varphi)} \parallel} \tag{8}$$

with x the original signal frame and $\parallel f \parallel^2 = \int f^2(t) dt$.

IV. Results

To evaluate the performance of the proposed audio coder, we compare it against a number of existing audio coders. The benchmark includes two audio coding standards - MP3

(ISO/IEC 11172-3 MPEG Layer 3) [18] and AAC (ISO/IEC 13818-7 Advanced Audio Coding) [19] codecs, and wavelets compression based approach. The Daubechies wavelet Db8 is used as mother wavelet. This function is orthogonal 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 coding tool. In general, this wavelet gives good audio coding results compared to other wavelets [8]. Test material are taken from the European broadcasting union Sound Quality Assessment Material (SQAM) CD. The audio files are gspi, harp, quar, song, trpt, classical, orchestra and castanet (percussive sound) and are sampled at 44.1 kHz. These eight tracks, illustrated in figure 5, are chosen to represent a variety in audio content. Compared to our previous approach [11], in the present work coding performances are analyzed using the BR, the Objective Difference Grade (ODG) and the Noise to Mask Ratio (NMR) [20]. The ODG represents the expected perceptual quality of the degraded signal if human subjects are used. This criterion ranges from 0 to -4 where 0 represents a signal with imperceptible distortion and -4 represents a signal with very annoying distortion [21]. This objectively measured parameter is calculated by perceptual evaluation of the audio quality algorithm specified in ITU BS.1387-1 [20]. The NMR is an objective measure of the perceptual quality of a compressed signal which measures the relative level of the quantization noise in comparison with the masking threshold [22]. Lower coding errors are indicated by larger negative values of NMR. It has been shown that NMR is useful tool in the development and comparison of perceptual coding schemes. In our previous work, we essentially focus on the quality of coding/decoding signal rather than on the BR [11]. In particular, it was not possible in [11] for comparison purpose to fix the $\mathtt{B}_s^r$ to 64 kb/s, so the coding was only compared to MP3 codec with a $\mathtt{B}_s^r$ set to 96kb/s. In the present work, since the order of IF model is variable, in addition to MP3 codec, we

also include AAC codec for comparison with $\mathtt{B}_s^r$ set to 64 kb/s. An interest of the proposed IF encoding strategy is to support variable BR coding. This is mainly due to the fact that the number of coefficients to be coded varies from one mode to another and this number is not subject to perceptual constraints. The frame length L is set to 512 and the overlap length between adjacent frames is set to 64.AR orders (p_k^l, q_k^l) are encoded with $m_o = 4$ bits. For the sake of simplicity, we used the uniform scalar quantization. For the IA component of each IMF, the order p (or p_k^l) is calculated based on a perceptual constraint [11]. More precisely, the order p_k^l is adjusted so that the PSD of the reconstruction error between the original IMF and the reconstructed one does not exceed the masking threshold of this IMF. According to our previous results the order $p_k^l = 9$ satisfies the PSD constraint and represents a good compromise between compression ratio and listening quality measured using subjective difference grade [11]. For IF component we used the partial autocorrelation plot of the associated mode to identify the order q_k^l of the AR model (Fig. 4).

Values of NMR, BR and ODG obtained with the HHT-coding, MP3, AAC and wavelets approaches are shown in Table II. A careful comparative examination of the values reported in this Table shows that the proposed coding outperforms the MP3 and wavelets approaches and on average performs better than the AAC codec in terms of ODG and NMR. The effectiveness of the proposed approach is mostly shown in the signals "gspi", "song", "trpt" and "castanet", where this last performs better than the AAC codec. When listening the decoded signal, the proposed approach yields overall better perceptual quality compared to the others techniques, mostly in the signals "gspi", "song", "trpt" and "castanet". Compared to wavelet approach and MP3 codec, for all signals both HHT-coding and AAC codec have ODG values between 0 (not perceptible) and -1 (not annoying). These results show that the performances coding of the proposed method are obtained without significant perceptual

distortion. In terms of NMR, it can be observed from Table II that on average larger negatives values are given by the proposed coding leading to lower coding errors and less audible noise. Overall, when compared to MP3, AAC and wavelet methods, there is a preference towards HHT approach for the eight tested audio signals.

Above results are obtained with initial phase of audio signals set to zero (Eq. (7)). The influence of this phase on reconstructed IMFs of each frame is analyzed through random phase initialization. For each audio signal, 100 frames are selected randomly and, for each one, initial phase is initialized randomly between 0 and 2π. This procedure repeated 20 times. These simulations are performed for the eight tracks and the corresponding error E_φ is calculated from Eq. (6). Results in Fig. 6 show that over all the analyzed signals E_φ does not exceed 6% showing that an arbitrary choice of initial phase has limited effect on the reconstructed IMFs. More important, we have seen that the perceptual quality remains high, despite this phase distorsion. However, initial phase could also be encoded for better IMFs recovery at the expense of slight bitrate increase.

Performances of HHT-coding is assessed on different audio signals (songs and instruments). Even these signals have different frequency contents, the obtained performances show that these signals are well-modeled or represented with a reduced number of IMFs (atoms), mainly due to the adaptive nature of the EMD. No prior assumptions have been made about these signals concerning the number of IMFs for their expansions and their codings. Further, the coding requires only two parameters, the frame size (L) and size of LEC sliding window (N). The reported results in terms of BR, NMR and ODG also show the interest to code both IA and IF components.

Since it is based on EMD, HHT-coding is a data driven approach. The number of IMFs per frame and the associated AR orders (p_k^l, q_k^l) is not fixed a priori. The amount of allocated

bits is also signal dependent. Performance of HHT-coding is evaluated using NMR and ODG criteria at BR fixed to 64kb/s. BR variations are supported by IF coding while good listening quality of IA coding is controlled by a psychoacoustic model. The association of both codings yields good results for BR$\geq$ 64kb/s. For BR$<$ 64kb/s quality degradation can be perceived. More precisely, audio quality deteriorates noticeably for BR $\leq$ 50kb/s. A solution to improve the quality of the decoded signals at very low BR is to partially reconstruct the frame signal with a set of selected IMFs controlled by the psychoacoustic model. We expect that the reduction of the number of coded IFMs per frame, will enable higher quality of decoded signals at low BR ($\leq$ 50kb/s). This idea is currently under investigation.

Performance of the HHT-coding depends on the quality of the sifting. Thresholding of the stopping criterion is set to $\epsilon = 0.25$, according to Huang et al. [9] who suggest to set it between 0.2 and 0.3. The coding has been tested with values of ϵ ranging from 0.2 to 0.3 and no changes have been noted in coding performance. For interpolation, the definition of IMF does not specify what is required for upper and lower envelopes, but that they pass through the maxima and minima of the signal respectively. Interpolations such as linear and polynomial tend to increase the required number of sifting iterations and to over-decompose signals by spreading out their components over adjacent modes. B-spline interpolation is commonly used to approximate upper and lower envelopes in EMD [27]. A recent study has shown that trigonometric interpolation is useful from an analytical point of view but involves more computational complexity than B-splines interpolation. Thus, the study in [28] does not recommend to use it in place of B-splines interpolation. Previous works motivate our choice for B-splines.

Based on results obtained for a variety of audio files, we believe that EMD-based coding can

be extremely promising for large classes of signals (not limited to audio signals) such as, for example, biomedical signals (ECG,...).

V. Conclusion

This paper presents a new audio coding. The scheme is based on the HHT and consequently is simple and fully data-driven method. Also, the bit allocation process is signal dependent. The obtained results in terms of ODG and NMR criteria show that, overall, the HHT-coding outperforms MP3 and AAC codecs, and wavelets approach. The effectiveness of the coding scheme is obtained especially for audio signals "gspi", "song", "trpt" and and "castanet", where HHT-coding performs better than AAC codec. Experimental results show that our scheme has good coding performances without significant perceptual distortion (ODG in the range $[-1, 0]$) and with lower coding errors (large negative values of NMR). These results show the interest to code both IA and IF components. Also, these findings show the interest to code the IF whether the instantaneous phase [11] to support variation BR, and demonstrate the potential of the EMD as a promising audio coding tool. Although experiments have already been carried out on different audio signals, future works should consider large classes of audio signals as well as varied experimental conditions such as different sampling rates or frame size for improving again the performances of the method. We also plan to find a strategy to encode efficiently the initial phase for better IMFs recovery and without increase of BR. Limited to BR greater than 64kb/s, also as future we plan to extend the HHT-coding for very lower BR while keeping a good listening quality of the audio signals. To quantify the stationarity of the frame, LEC method is used due to its simple use. It would be interesting to check if there is enough stationarity in the frame using a time-frequency method such as spectrogram [23] or Wigner-Ville distribution [24]. Also, instead of EMD, it would be interesting to use the ensemble EMD [25] or complete ensemble

EMD [26] to investigate the performances of the proposed coding.

Acknowledgments

The authors are grateful to the anonymous reviewer for supplying useful comments and pointing out the proof-of-concept nature of the paper.

References

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

[2] P. Noll, "MPEG digital audio coding," IEEE Sig. Process. Magazine, vol. 14, pp. 59-81, 1997.

[3] 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, pp. 780-792, 1994.;

[4] G. Stoll, G. Theile, S. Nielsen, A. Silze, M. Link, R. Sedlmeyer and A. Brefort, "Extension of ISO/MPEG-audio layer II to multi-channel coding. The future standard for broadcasting, telecommunication, and multimedia applications," *Conv. Aud. Eng. Soc.,* 1993.

[5] J.B. Rault, P. Philippe and M. Lever, "MUSICAM (ISO/MPEG audio) very low BR coding at reduced sampling frequency," *Conv. Aud. Eng. Soc.,* 1993.

[6] 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.

[7] P. Srinivasan and L. H. Jamieson, "High quality audio compression using an adaptive wavelet packet decomposition and psychoacoustic modeling," *IEEE Trans. Sig. Process.,* vol. 46, pp. 1085-1093, 1998.

[8] P.R. Deshmukh, "Multiwavelet decomposition for audio coding," *IE(I) Journal-ET,* vol. 11, pp. 38-41, 2006.

[9] N.E. Huang et al., "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Royal Society,* vol. 454, pp. 903-995, 1998.

[10] K. Khaldi, A.O. Boudraa, M. Turki, Th. Chonavel and I. Samaali, "Audio encoding based on the empirical mode decomposition," *Proc. EUSIPCO,* pp. 1-5, 2009.

[11] K. Khaldi, A.O. Boudraa, B. Torrésani, Th. Chonavel and M. Turki, "Audio encoding using Huang and Hilbert transforms," *Proc. ISCCSP,* pp. 1-5, 2010.

[12] A. Spanias, T. Painter and V. Atti, *Audio Signal Processing and Coding,* Wiley-Interscience, 464 pages, 2007.

[13] P. Flandrin, G. Rilling and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Sig. Proc. Lett.,* vol. 11, pp. 112-114, 2004.

[14] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech. Sig. Proc.,* vol. 34, pp. 744-754, 1986.

[15] Ph. Depalle, G. Garcia and X. Rodet, "Analysis of sound for additive synthesis: tracking of partials using Hidden Markov model," *Proc. ICMC,* vol. 34, pp. 94-97, 1993.

[16] G. Gonon, S. Montresor and M. Baudry, "Improved entropic gain and adaptive time-frequency segmentation. Application to audio coding," *EUROSPEECH,* vol. 4, pp. 2661-2664, 2001.

[17] T. Welch, "A technique for high-performance data compression," Computer, vol. 17, pp. 8-19, 1984.

[18] 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.

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

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

[21] D. Campbell, E. Jones and M. Glavin, "Audio quality assessment techniques-A review, and recent developments," *Signal Processing,* vol. 89, pp. 1489-1500, 2009.

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

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

[24] W. Martin and P. Flandrin, "Detection of changes of signal structure by using the Wigner-Ville spectrum," *Signal Proc.,* vol. 8, pp. 215-233, 1985.

[25] Z. Wu and N.E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Advances in Adaptive Data Analysis,* vol. 1, pp. 1-41, 2009.

[26] M.E. Torres, M.A. Colominas, G. Schlotthauer and P. Flandrin, "A complete Ensemble Empirical Mode decomposition with adaptive noise," *IEEE ICASSP*, pp. 4144-4147, 2011.

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

[28] S.D. Hwaley, Les E. Atlas and H.J. Chizeck, "Some properties of an empirical mode type signal decomposition algorithm," *IEEE Sig. Process. Lett.,* vol. 17, no. 1, pp. 24-27, 2010.

TABLE I
ORDER OF AR MODEL OF IF COMPONENTS (FIG. 4).

IMF	1	2	3	4	5	6
Order of AR model	11	7	8	7	8	4

TABLE II
COMPRESSION RESULTS OF AUDIO SIGNALS (GSPI, HARP, QUAR, SONG, TRPT, CLASSICAL, ORCHESTRA AND CASTANET) USING HHT-CODING, AAC, MP3 AND WAVELETS APPROACHES.

	Signal	gspi	harp	quar	song	trpt	classical	orchestra	castanet
HHT coding	BR [kb/s]	64	64	64	64	64	64	64	64
	NMR	**-4.65**	-5.79	**-5.47**	-4.21	**-6.80**	**-6.22**	-4.87	-8.32
	ODG	**-0.74**	-0.73	-0.75	**-0.72**	**-0.84**	-0.91	-0.86	**-0.67**
AAC	BR [kb/s]	64	64	64	64	64	64	64	64
	NMR	-3.43	**-6.46**	-4.78	**-4.23**	-6.15	-6.02	**-7.48**	-8.32
	ODG	-0.85	-0.73	-0.75	-0.89	-0.88	**-0.83**	**-0.74**	-0.70
MP3	BR [kb/s]	64	64	64	64	64	64	64	64
	NMR	1.42	1.21	1.27	1.23	2.68	1.17	1.34	1.68
	ODG	-1.12	-1.87	-1.91	-1.09	-1.27	-1.05	-0.95	-1.12
Wavelets	BR [kb/s]	65	67	64	65	66	67	66	64
	NMR	-2.30	-3.67	1.64	-3.40	-1.35	-3.01	-4.27	-2.32
	ODG	-0.86	-1.27	-1.74	-0.98	-0.97	-1.02	-0.97	-0.94

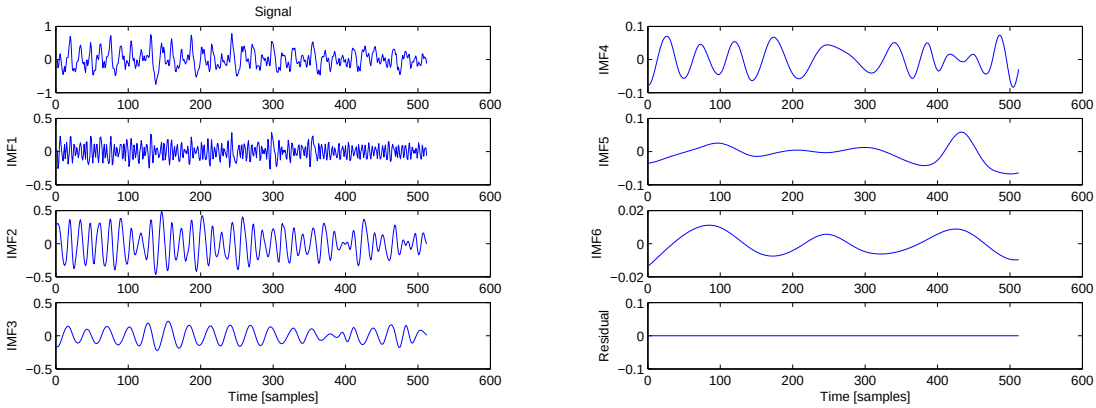

Fig. 1. Decomposition of an audio frame by EMD.

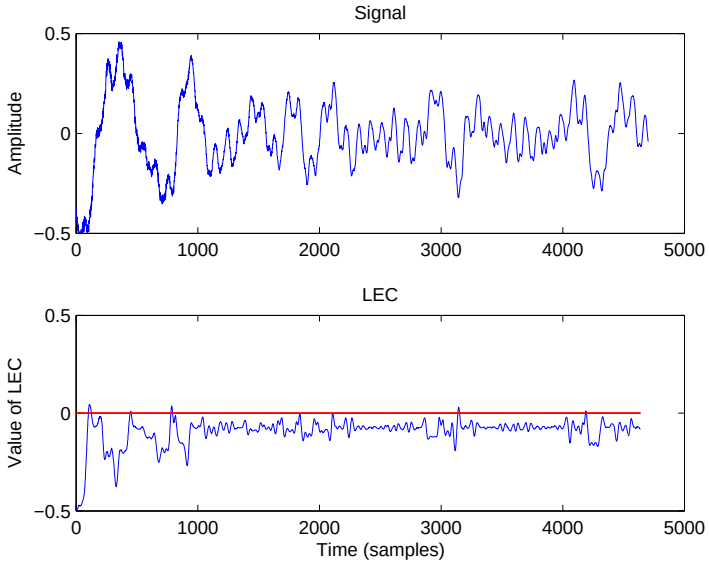

Fig. 2. LEC variations of an audio frame.

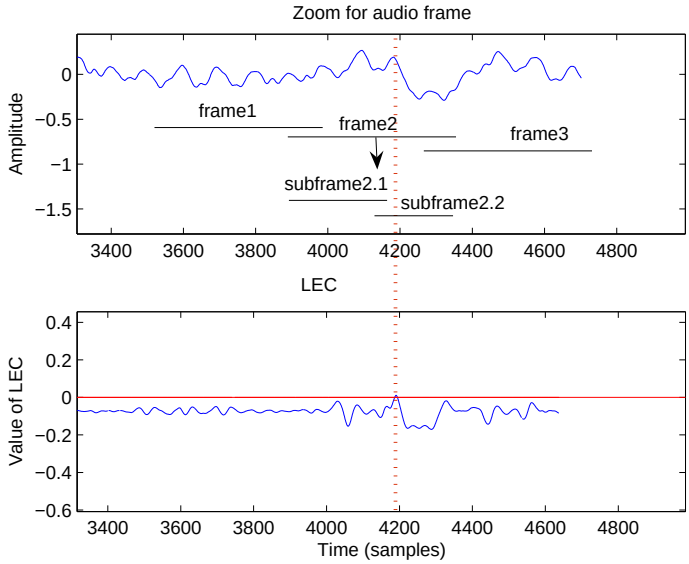

Fig. 3. Example of segmentation of an audio frame.

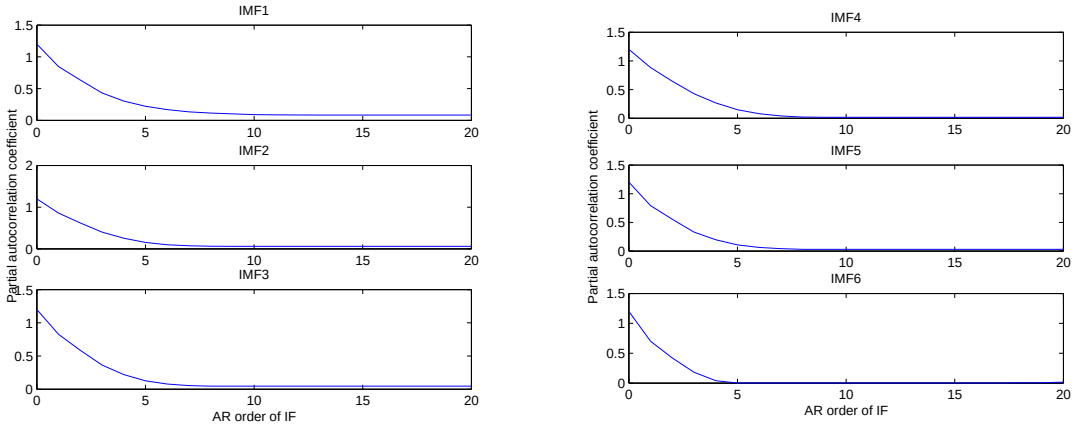

Fig. 4. Partial autocorrelation coefficient of IF of the IMFs extracted from a frame of audio signal (Fig. 1).

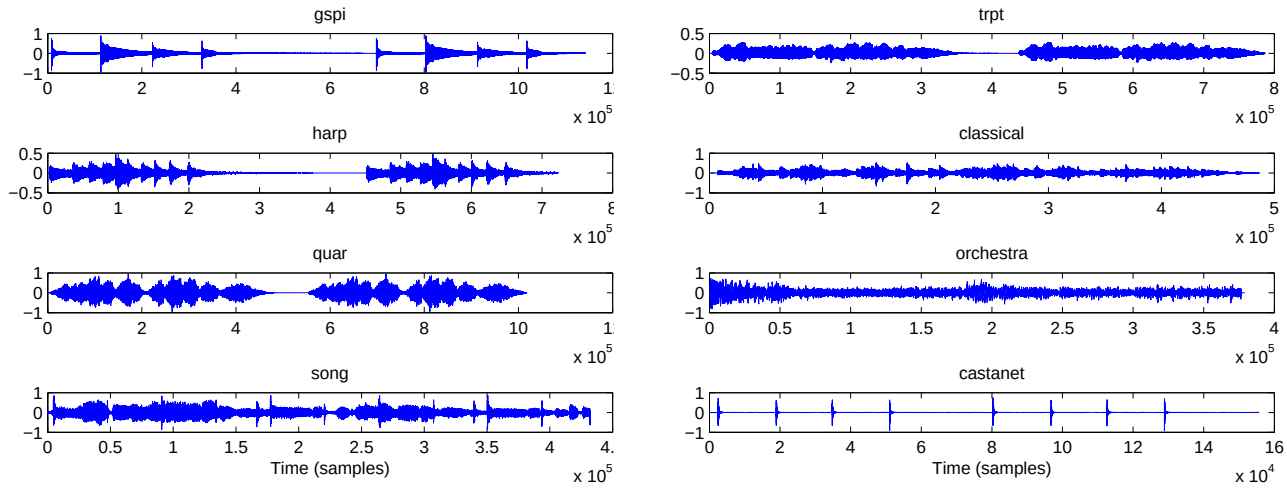

Fig. 5. Original audio signals (gspi, harp, quar, song, trpt, classical, orchestra and castanet).

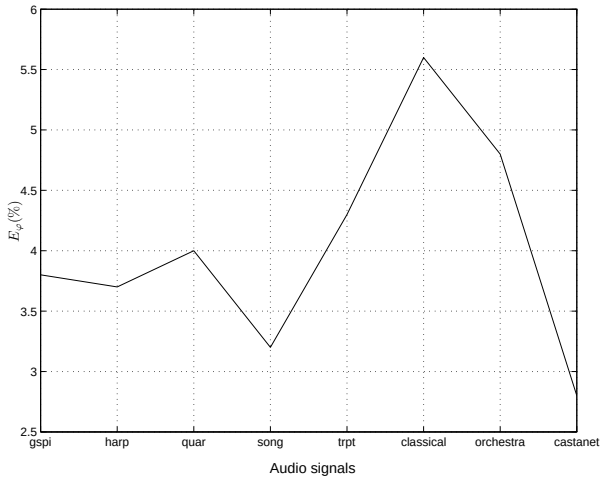

Fig. 6. Effect of initial phase on IMFs recovery.