Absolute optimal time - frequency basis - a research tool

1997 J. Phys. A: Math. Gen. 30 7133

(http://iopscience.iop.org/0305-4470/30/20/016)

J. Phys. A: Math. Gen. **30** (1997) 7133–7146. Printed in the UK PII: S0305-4470(97)83506-8

Absolute optimal time–frequency basis—a research tool

Alin Cârsteanu, Victor B Sapozhnikov, V Venugopal and Efi Foufoula-Georgiou

St Anthony Falls Laboratory, Mississippi River at 3rd Ave SE, University of Minnesota, Minneapolis, MN 55414, USA

Received 17 April 1997

Abstract. The paper presents a method for finding the absolute best basis out of the library of bases offered by the wavelet packet decomposition of a discrete signal. Data-adaptive optimality is achieved with respect to an objective function, e.g. minimizing entropy, and concerns the choice of the Heisenberg rectangles tiling the time–frequency domain over which the energy of the signal is distributed. It is also shown how optimizing a concave objective function is equivalent to concentrating maximal energy into a few basis elements. Signal-adaptive basis selection algorithms currently in use do not generally find the absolute best basis, and moreover have an asymmetric time–frequency adaptivity—although a complete wavepacket decomposition comprises a symmetric set of tilings with respect to time and frequency. The higher adaptivity in frequency than in time can lead to ignoring frequencies that exist over short time intervals (short as compared to the length of the whole signal, not to the period corresponding to these frequencies). Revealing short-lived frequencies to the investigator can bring up important features of the studied process, such as the presence of coherent ('persistent') structures in a time series.

1. Introduction

1.1. Information representation in the time–frequency plane

The development of time–frequency representations of signals (see e.g. [1]) is an area of continuing interest in many engineering and scientific disciplines, for the purpose of investigating and/or compressing observed signals or processes. While the latter goal is reached by using non-redundant representations, the former can also be pursued by redundant (possibly highly redundant, in the limit continuous) representations. Continuous representations give the researcher the opportunity to observe fine-structure features such as branching and edges, while non-redundant representations may be optimized (e.g. least-entropy) to capture much of the information about the signal in a few numbers. In this way, they make it possible to unravel underlying laws of the studied phenomenon from the positions in the time–frequency plane and the shapes of the Heisenberg tiles which contain most of the signal energy. If orthogonality is preserved within a non-redundant representation (a basis), then the areas of the Heisenberg tiles are non-overlapping, i.e. each basis element represents, to a good extent, a distinct area of the time–frequency plane. These orthogonal bases are the focus of the present article, which shows how to tile the time–frequency plane in such a fashion so as to capture the most possible information in a few tiles, and to gain a better understanding of the process from the way the time–frequency plane is being optimally tiled. To that end, we present a straightforward algorithm that we

0305-4470/97/207133+14$19.50

have developed for optimizing the tiling with respect to an objective function that is additive over the elements of a basis.

1.2. Fundamental tilings of the time–frequency plane

In a general sense, frequency can be defined as the number of repetitions of a certain pattern per unit time. A frequency representation of a signal can be best understood by an analogy to time representation, i.e. while in the latter, intensity is considered at different points in time, in the former, intensity is considered at different frequencies. In other words, the intensity corresponding to a particular frequency is representative of the contribution of that frequency to the signal. This intensity can be shown to result from the convolution of the signal with the (chosen) pattern of that frequency. For instance, a classic example of a frequency representation is the Fourier transform, wherein the pattern is a sinusoid.

The frequency representation that we have mentioned 'looks' at the signal as a whole, i.e. globally. Such representations would be useful for processes where the frequency content does not change over time. However, the frequency content of many natural processes changes considerably over time and the aforementioned representation(s) fall short in assisting one to characterize these changes. A solution to this drawback is representations that analyse parts of the signal separately. Such representations are called short-time frequency representations.

A 'packet' of all such short-time frequency representations can be built for a given signal and a chosen analysing pattern. Using, for example, the Haar wavelet as our pattern (which is the simplest pattern available when smoothness is not necessary, and moreover, it does not require the use of an additional, different wavelet at the boundaries, to preserve orthogonality [2]), we build the Haar wavelet packet as outlined by Wickerhauser [3], and explained below. If $\boldsymbol{r} = (r_1, r_2, \ldots, r_n)$ denotes a discrete sequence of $n = 2^k$ values, we define the 'high-frequency' (or differencing) operator $H(r_{2i-1}, r_{2i}) = (r_{2i-1} - r_{2i})/\sqrt{2}$, and the 'low-frequency' (or averaging) operator $L(r_{2i-1}, r_{2i}) = (r_{2i-1} + r_{2i})/\sqrt{2}$, where $i = 1, \ldots, n/2$. Applying either one of the operators to the sequence $\boldsymbol{r}$ in a non-overlapping manner, we obtain a sequence of $n/2$ elements, symbolically denoted $H\boldsymbol{r}$ or $L\boldsymbol{r}$, respectively. Note that the recursive, alternate application of the two operators to a data series is equivalent to convolving subsequences of the signal with the Haar wavelet at different frequencies [3]. Importantly, the construction *rule* for the wavelet packet is independent of the particular wavelet ('pattern') that is used [3]. We obtain the tree shown in figure 1, which illustrates the procedure for an 8-value series.

Note that the horizontal levels in the wavelet packet are (top-to-bottom): the signal itself (the 'temporal representation' or 'standard basis'), the discrete short-time frequency representations (DSTFRs), and finally the frequency representation. Note that for a time series containing $n = 2^k$ data points, there are exactly $\log_2(n) - 1 = k - 1$ different DSTFRs. Together with the temporal representation and the frequency representation, they make up the $\log_2(n) + 1 = k + 1$ possible tilings of the time–frequency domain with n identical, similarly-positioned rectangles. These are called here the fundamental representations. Moreover,

<table>
<tr><td colspan="8">r</td></tr>
<tr><td colspan="4">Lr</td><td colspan="4">Hr</td></tr>
<tr><td colspan="2">LLr</td><td colspan="2">HLr</td><td colspan="2">HHr</td><td colspan="2">LHr</td></tr>
<tr><td>$LLLr$</td><td>$HLLr$</td><td>$HHLr$</td><td>$LHLr$</td><td>$LHHr$</td><td>$HHHr$</td><td>$HLHr$</td><td>$LLHr$</td></tr>
</table>

Figure 1. Wavelet packet.

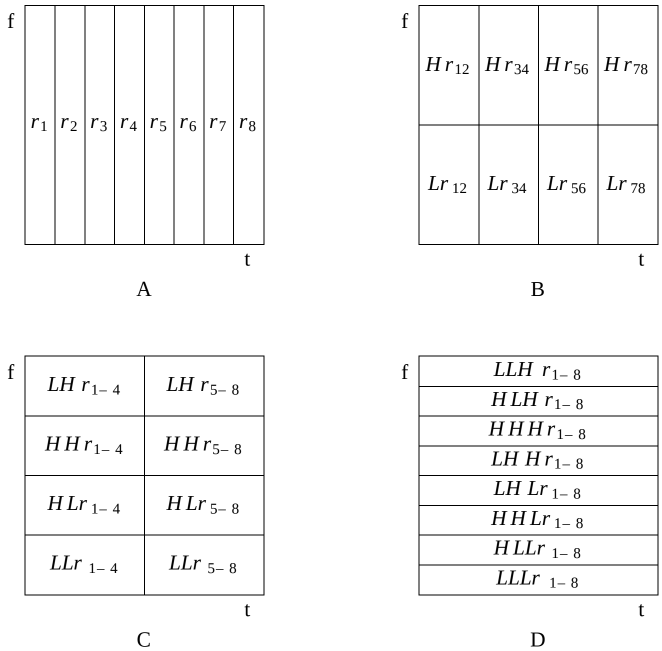

Figure 2. Tilings corresponding to the wavelet packet fundamental representations (Haar-wavelet indexing) of a time series containing eight values.

these $k+1$ tilings exhaust all shapes of rectangles that appear in tilings of the time–frequency domain of such a diadic decomposition of a time series.

Symbolic diagrams of these representations can be drawn in the time–frequency plane, as shown in figure 2 (for an 8-value series, again). The tiles in the diagrams symbolize the location of the respective wavelet-packet coefficients, in the sense of associating them with an area in the time–frequency plane. The tiles are significant in rendering the ratios of frequency resolutions as the ratios of their heights and the ratios of time resolutions as the ratios of their widths (due to the Heisenberg principle), as well as their relative positioning in time and in frequency. However, their absolute sizing and positioning do not have a precise meaning by themselves, but stand in relation with the time–frequency localization of the particular wavelet that is used [2].

In using time–frequency representations of a process as a research tool, it is important that the Heisenberg tiles cover the time–frequency plane without gaps (and, consequently, without overlaps), in order to represent each area of the time–frequency plane in the basis. Non-overlapping is also equivalent (for the Haar wavelet used herein) to the orthogonality property of the basis. Non-overlapping implies orthogonality since two non-overlapping rectangles have non-overlapping projections onto at least one axis. Without loss of generality we assume it is the time axis. According to the construction rule in equations (1), the basis element corresponding to any given tile can be obtained from the signal elements whose tiles project onto the same time interval as that given tile. Therefore, the signal elements involved in representing the two non-overlapping tiles are disjoint, and their scalar product is zero. Conversely, two overlapping tiles have overlapping projections on both axes and hence are not orthogonal.

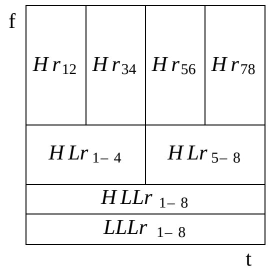

Figure 3. Wavelet tiling.

2. Background on some classical time–frequency representations and algorithms

2.1. The wavelet representation

In essence, the wavelet representation is a first attempt to adapt a combination of elements of fundamental representations to features of the data. It stems from the fact that very often, in natural signals, several frequencies present in the signal have bandwidth proportional to the values of those frequencies, i.e. comparable energy content is found within geometrically progressing frequency bands [4, 5]. This is equivalent to saying that at each frequency, the time resolution (inverse 'scale') is made proportional to the period of the signal, that is, the length of the signal involved in calculating one wavelet-basis coefficient is equal to one pattern length at every scale.

Consider the 8-values example in figure 2 to see how one can go to a wavelet representation from the fundamental representations. The choice made at the first step is the top half (i.e. $Hr_{12} \dots Hr_{78}$) from B; the next step is the choice of the upper blocks in the lower half (i.e. HLr_{1-4} and HLr_{5-8}) from C and the final step involves the selection of the bottom most two elements (i.e. $HLLr_{1-8}$ and $LLLr_{1-8}$) from D. The so obtained representation is called the wavelet representation and is illustrated in figure 3. It is worth mentioning that such a selection leads to a constant stepsize in log-frequency.

2.2. The 'classical' single-tree algorithm

While the wavelet representation shown in figure 3 was historically the first one to result in combining elements of different D(ST)FRs in a *fixed* tiling (according to the criterion of geometric progression of frequencies), the first procedures to perform an *adaptive* selection of elements from the fundamental representations came with the tree algorithms [3, 6]. In order to capture the maximal energy in any given number of basis elements (obtain 'maximal contrast'), minimal entropy is one appropriate objective function in a context where the L^2-norm (energy) of the basis is preserved. This choice is motivated by the equivalence of negative entropy to information [7]. Thus, in choosing and combining tiles of the described fundamental representations, we seek herein a representation with minimal entropy. However, in the appendix it is shown that any concave objective function produces the same ordering within a set of normalized bases as the energy contrast ordering, if the latter exists.

The algorithm that allows us to pick an adaptive representation based on an entropy objective function (or any concave objective function whose value over a set of tiles is the sum of the values for each tile) works [3] as follows. Starting at the lowest level, each pair of adjacent blocks is compared with the block immediately above, in terms of entropy.

The one satisfying our minimal entropy criterion is 'promoted' (chosen). Then, this step is repeated $k = \log_2(n)$ times, up to the level of the signal. The resulting representation is optimal *within a certain pool*, and the algorithm has a complexity of order $n \log(n)$.

To express the entropy, we define the normalized energy of the ith sample of a time series as

$$e_i = \frac{r_i^2}{\sum_{j=1}^{n} r_j^2}$$

and similarly at each (horizontal) level of the wavelet packet (WP) tree, as the square of the intensity corresponding to the ith tile, normalized by the sum of squares of all intensities at that level (i.e. the intensities corresponding to all tiles of that fundamental representation). It is worth mentioning that normalization is superfluous if energy-preserving operators, such as in equations (1), are used. From the energy e_i we define the entropy of the ith tile as

$$s_i = -e_i \ln(e_i).$$

Disadvantages of the WP single-tree algorithm. Every block in the WP tree involves a time span that is equal to the length of the signal. Since the process of selection involves only whole blocks, the time resolution of our representation of choice is affected: although different intensities do indeed show along the time axis, *the split itself* (i.e. the shapes of the rectangles in the time–frequency tiling) *is not time-adaptive*. Two negative consequences are obvious: (a) in an immediate perspective, we do not reach the absolute minimal entropy, since we only look at a subset of the tiles constructed from elements of the fundamental representations; and (b) in a wider perspective, we do not achieve symmetry between time and frequency. As will be shown later, this asymmetry of the tiling can obscure important features of a signal.

2.3. The double-tree algorithm

To overcome the above-stated shortcomings, Herley *et al* [8, 2] as well as Ramchandran and Vetterli [9] have elaborated new algorithms, of which the most performant is the double-tree algorithm. Its complexity is of order $n \log^2(n)$, with approximately the same proportionality constant as the original WP-tree algorithm. Essentially, the algorithm amounts to running the aforementioned WP-tree algorithm on each of the fundamental representations. The double-tree algorithm also fails, in certain cases, to find the absolute best basis, since it does not exhaust all tiling comparisons. For some synthetic signals this non-optimal selection of bases can be shown to result in significant differences in entropy, although in most real-life cases (the rainfall series we studied) these differences had relative values of only a few percent. For such purposes as data compression and/or transmission, this algorithm may be appropriate, especially given its speed. For research purposes however, one may want to have an algorithm that really finds the optimal tiling. Remarkably, for the rainfall intensity process, which is positive and consequently has a strong low-frequency component, the double-tree algorithm renders mostly the *same* decomposition as the single-tree algorithm. For illustration purposes, figure 4 shows the entropy-minimizing tilings for the sequence $(5, 0, 2, 4, 8, 6, 1, 0)$ as rendered by the single-tree algorithm, the double-tree algorithm, as well as by the algorithm presented in the next section, which finds the absolute best basis.

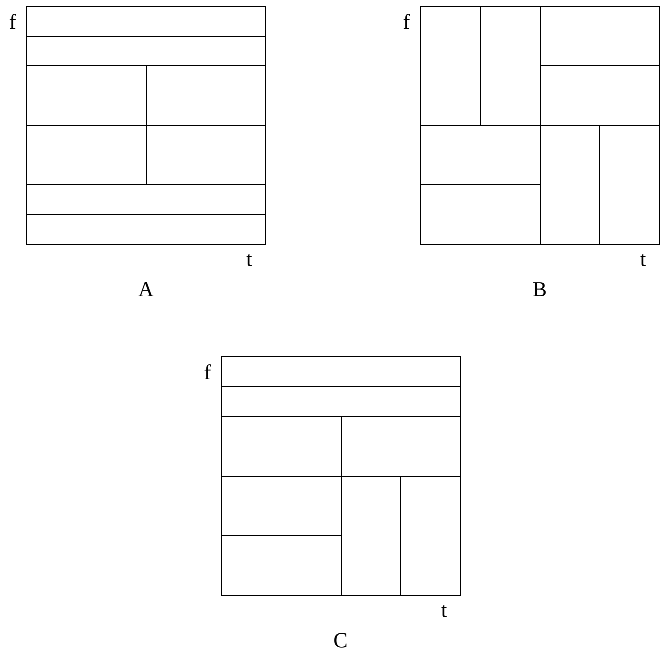

Figure 4. Entropy-minimizing time–frequency tilings for the sequence $(5, 0, 2, 4, 8, 6, 1, 0)$, as rendered by the single-tree algorithm (A), the double-tree algorithm (B), and the optimal-basis algorithm presented in section 3 (C). The respective entropies of the normalized bases are approximately 1.071, 0.967, and 0.95.

2.4. *Reconstruction*

Note that what makes these algorithms work is the fact that any pair of adjacent blocks in the WP-tree is interchangeable with the block above. This is since

$$\left.\begin{array}{l} L(r_{2i-1}, r_{2i}) = (r_{2i-1} + r_{2i})/\sqrt{2} \\ H(r_{2i-1}, r_{2i}) = (r_{2i-1} - r_{2i})/\sqrt{2} \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} r_{2i-1} = (L(r_{2i-1}, r_{2i}) + H(r_{2i-1}, r_{2i}))/\sqrt{2} \\ r_{2i} = (L(r_{2i-1}, r_{2i}) - H(r_{2i-1}, r_{2i}))/\sqrt{2}. \end{array}\right. \qquad (1)$$

This inductively shows that each fundamental representation is a basis in $\boldsymbol{r}$, with elements being linearly-independent combinations of $(r_1, r_2, \ldots, r_n)$. The above equation has two other immediate consequences: (a) in the time–frequency tiling diagram, any rectangles that fill the same area are interchangeable while transforming a basis into another basis; and (b) a signal-reconstruction algorithm with complexity of order $n \log(n)$ can be written for an arbitrary representation (built from elements of the fundamental representations), by simply replacing pairs of elements from the lower levels with their L-and-H inverse operators as shown above.

3. An algorithm to choose the absolute optimal basis

For research purposes, it is not only the value of the basis element associated with a certain tile in the time–frequency plane that carries information. *The tiling shapes themselves*, since they are adaptive, can capture highly localized phenomena, and show the time–frequency localization of these phenomena. It is therefore our goal to both overcome the splitting restrictions of the 'classical' tree algorithm (which needs to keep the same tiling throughout

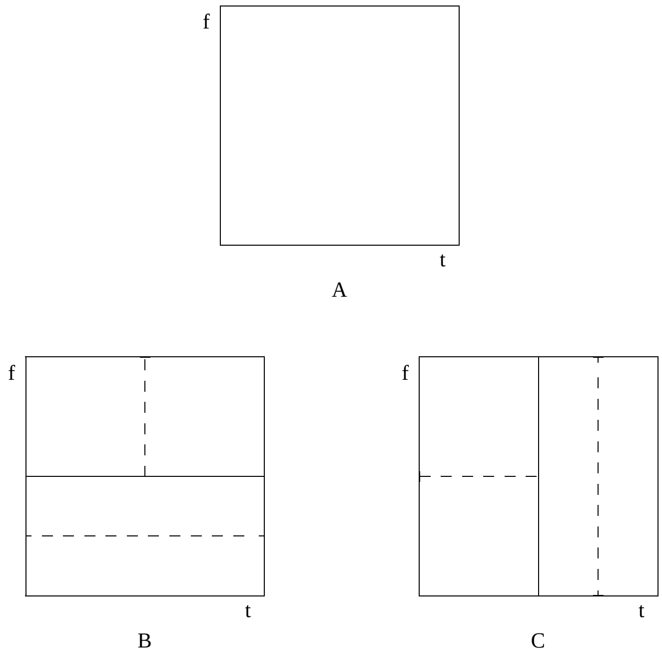

Figure 5. Illustration of the best-basis tiling algorithm.

time), and also to completely optimize the adaptive tiling, something that both the single-tree and the double-tree algorithm fail to do. Another issue of concern is the lack of shift-invariance of tree algorithms. Since in the analysis of a natural process there often is no clearly defined moment of incipience, we would like to have a tool whose results do not drastically vary with a time shift.

The idea of the proposed algorithm is the following: any diadic non-overlapping time–frequency tiling, and therefore the optimal representation, has to be split in half either by time, or by frequency, or both. Therefore, e.g., our initial 8-point time–frequency square (figure 5(A)) has to be halved as in either figure 5(B) or figure 5(C), full lines. We decide for the optimal tiling by further applying the above idea iteratively (e.g. the subsequent halving shown by broken lines in figures 5(B) and (C), and so forth) down to the elementary level, where each rectangle corresponds to exactly one DSTFR tile. Note that naturally, as we go back the decision path, at each level the existing rectangles contain the optimal tiling structure that can be achieved within each one of these rectangles. The algorithm takes $k = \log_2(n)$ steps for a discrete sequence of $n = 2^k$ values. Since at each step 2 sums, from 2 rectangles each, are to be compared in terms of the objective function, a complexity of order $4^k = n^2$ results. Note here that the number of possible tilings grows exponentially with n. (This can be shown by noting that the number of possible tilings of the time–frequency square with n tiles drawn from the fundamental representations can be written, from considerations of symmetry, as $T_n = 2T_{n/2}^2 - T_{n/4}^4$, with $T_1 = 1$ and $T_2 = 2$. We have from here that $\exists b$ such that $\infty > \lim_{n\to\infty} T_n/b^n > 0$, with $b = \Pi_{i=1}^{\infty} a_i^{-1/2} \approx 1.844\,547\,57$, $a_i = (2 - a_{i-1}^2)^{-1}$, $a_0 = 0$. Also, $\lim_{n\to\infty} T_n/b^n = \lim_{i\to\infty} a_i = (\sqrt{5} - 1)/2$, the inverse of the golden ratio.)

The algorithm completely optimizes the tiling of the time–frequency plane with respect to a given objective function (e.g. entropy), using tiles from the fundamental

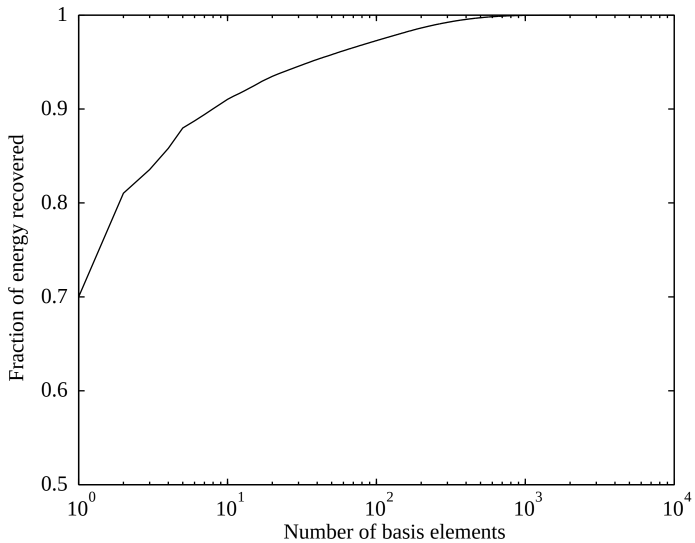

Figure 6. Fraction of energy recovered by the optimal WP basis elements (in decreasing order of magnitude) of the Iowa City rainfall event of 12 April 1991.

representations for which tile intensities are obtained with a given pattern function (e.g. Haar wavelet).

Proof. Halving p times the time–frequency square of a signal of length $n = 2^k$, $k > p \geqslant 0$, we obtain a rectangle that contains $n/2^p = 2^{k-p}$ tiles. Such a rectangle can contain tiles over its whole length, or over its whole width, or none of these, but not both (or otherwise they would overlap); as a consequence, it can always be halved either in frequency, or in time, or both. Or, in other words, the union of the sets of tilings of the two vertical halves of the rectangle and the sets of tilings of the two horizontal halves of the rectangle equals the set of tilings of the rectangle. This in turn implies that if we can find the optimal tiling of each of the two pairs of two halves, we can also decide which is the optimal tiling of the original rectangle. Since (a) p was chosen arbitrarily and (b) for $p = k - 1$ each half is one tile, the optimal tiling being decidable from directly computing the objective function, we conclude by induction that the halving algorithm described here finds the optimal tiling of the time–frequency square of the n-values signal.

Figure 6 shows how much of the energy of a rainfall time series is recovered by the elements (in decreasing order of magnitude) of the optimal representation.

The algorithm achieves the following:

(1) As it exhausts all combinations of ‘fundamental tiles’, the algorithm reaches the absolute optimal tiling with respect to the chosen objective function. Though the differences in the entropies of competing tilings might be small, the tilings themselves may be very different, and the truly-optimal tiling may depict, in the signal, patterns which might be of physical importance. A comparison between a ‘classical’ wavelet packet decomposition and the ‘best’ representation of the first 256 data points of the Iowa City 5-second rainfall time series from 12 April 1991 (complete time series shown in figure 7, and first 256 samples in figure 8) is shown in the top row of figure 9.

(2) To further see how the algorithm performs, let us superimpose a weak signal of

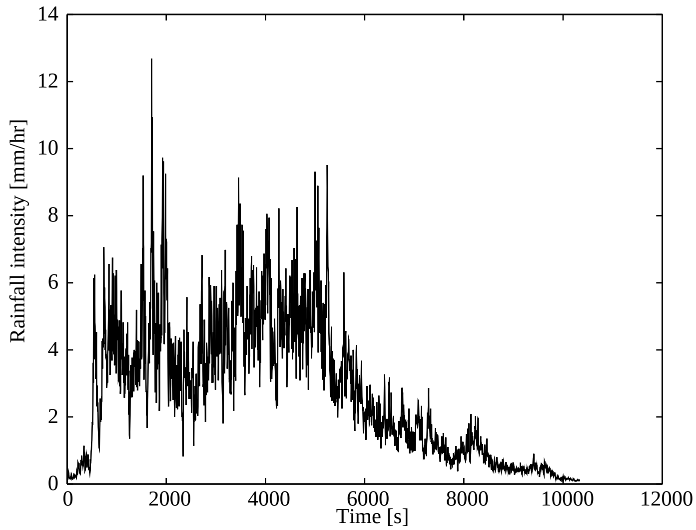

Figure 7. The Iowa City rainfall event of 12 April 1991.

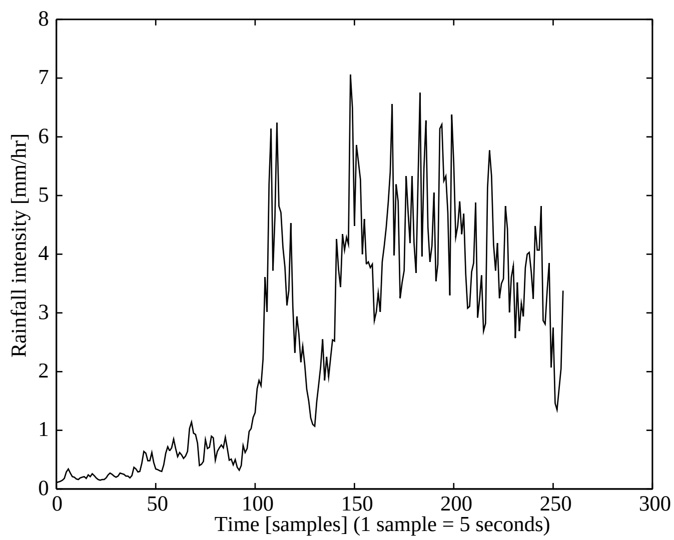

Figure 8. The first 256 samples (≈21 min) of the Iowa City rainfall event of 12 April 1991.

frequency $192 * (2 * 256 * 5\ \mathrm{s})^{-1}$ and amplitude 0.1 mm hr^{-1} on the first 16 data points. Note that the intensity of the superimposed signal is much lower than the average intensity of the rainfall, which is 2.48 mm hr^{-1}. The optimal representation shows that it is 'aware' of the superimposed signal by changing the tiling, as opposed to the single-tree algorithm, which only changes the intensities, but not the tiling in the area (see figure 9, second row). The amplitude of the superimposed signal required for the optimal representation to change the tiling is lower than the one for the single-tree algorithm.

(3) As a stronger superimposed signal (amplitude of 0.2 mm hr^{-1}) is used, the single-tree algorithm finally changes its tiling, but it has to do so throughout the whole time axis,

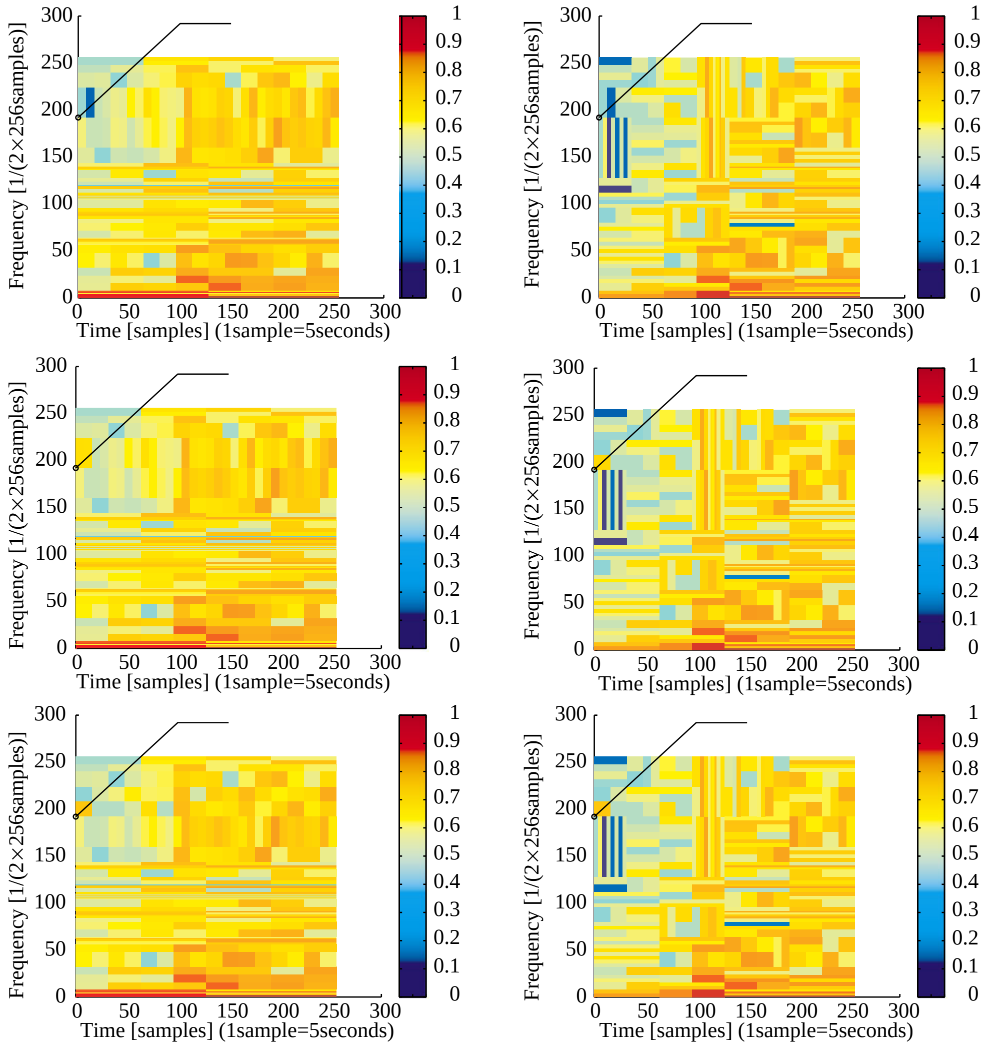

Figure 9. Classical' (left) versus optimal (right) wavelet packet representation of the first 256 data points (1280 s) of the Iowa City rainfall event of 12 April 1991, shown in figure 8 (top row; average rainfall intensity: 2.48 mm hr^{-1}); same, with a superimposed signal of frequency $192 * (2 * 256 * 5\ s)^{-1}$, as indicated by the pointer, and intensity 0.1 mm hr^{-1} for the first 80 s (second row); same, with intensity of the superimposed signal equal to 0.2 mm hr^{-1} (third row).

whereas the optimal representation is able to 'locally' depict the perturbation with respect to both time and frequency (see figure 9, bottom row).

(4) It is the case with the 'classical' tree decomposition that due to its tree structure, time-shift sensitivity is much higher than frequency-shift sensitivity. That makes its use problematic for research purposes, since the search for patterns in a signal should be as robust as possible with respect to the positioning of the analysis window, for which we generally have no *a priori* criteria. The optimal representations for two different, but overlapping segments of the Iowa City 5-second rainfall time series from 3 October 1990 (shown in figure 10) is displayed in figure 11. Clearly, the main features have been preserved

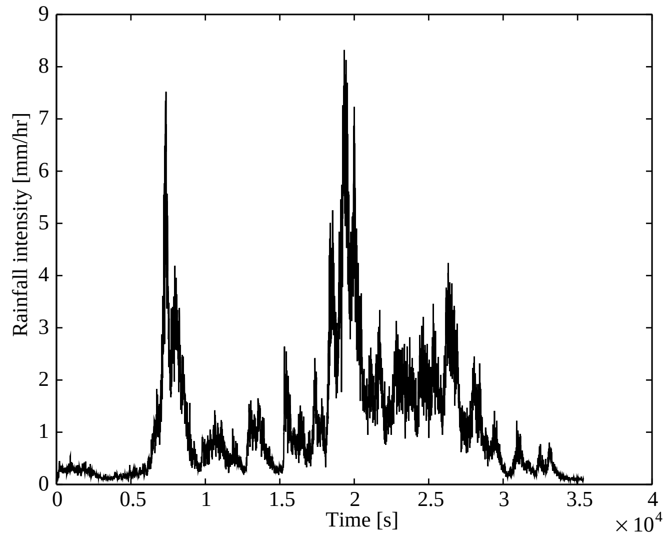

Figure 10. The Iowa City rainfall event of 3 October 1990.

in the overlapping zone. The same shifted signals are also represented using a single-tree decomposition ('classical' tree), and significant artificial differences show.

(5) The algorithm also exhibits time–frequency symmetry, overcoming the intrinsic disadvantage of all tree algorithms, namely their time–frequency asymmetry. Since in the presented algorithm the search takes place symmetrically in time and frequency, the results of that symmetry carry over to the optimal representation. As an illustration, a proper optimal representation of white noise should be statistically symmetric in the time–frequency domain (since both the time representation and the frequency representation are white noise). However, as is obvious from figure 12, the 'classical' tree algorithm gives a misleading result, i.e. the decomposition is strongly asymmetric, whereas the optimal basis reveals the actual time–frequency symmetry of this signal.

4. Conclusions

We have presented a method that enables one to find the absolute best basis from within a wavepacket decomposition of a signal, i.e. the basis that optimizes the value of the chosen objective function over all possible bases that can be constructed from that wavepacket decomposition. It provides a tool for revealing subtle features of a signal, which could be impossible to detect with the existing techniques. Although the chosen objective function in our article is entropy, we show in the appendix that all concave objective functions will yield the same results, i.e. will choose the same best basis that achieves maximal contrast (maximal energy in any number of highest basis elements), if such a maximal contrast is at all possible.

Acknowledgments

The authors would like to thank the Minnesota SuperComputer Institute for its continuous and generous support in computing and graphic resources. This research was partially

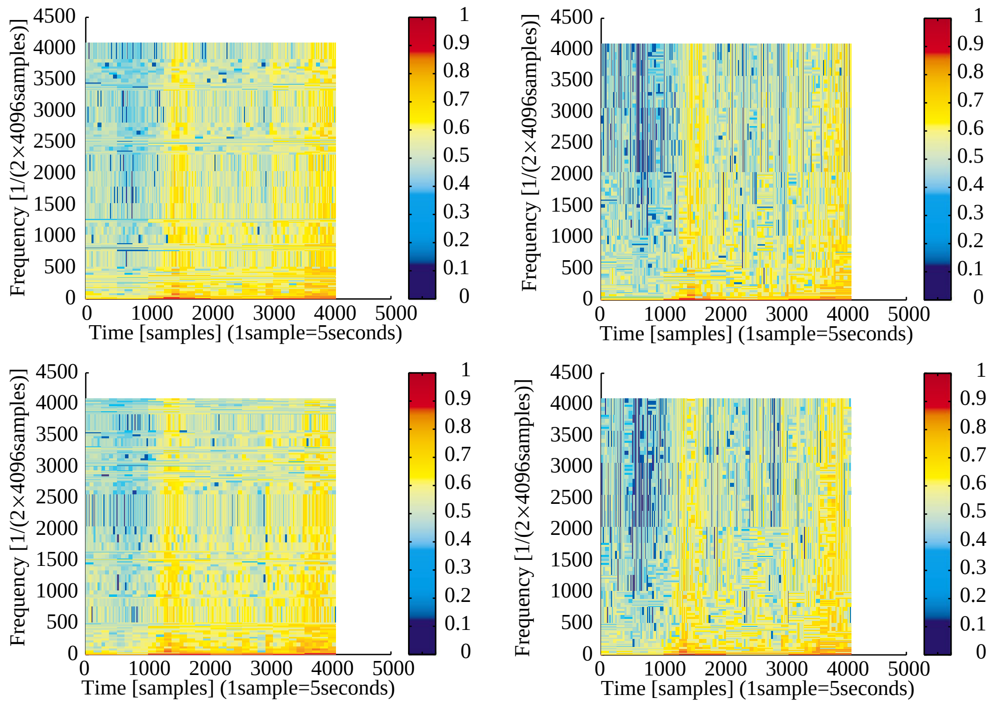

Figure 11. 'Classical' (left) versus optimal (right) WP representation of the first 4096 data points (20 480 s) of the Iowa City rainfall event of 3 October 1990 (top) and of the data points 101 to 4196 from the same event (bottom). The optimal tiling is very close to being shift-invariant, while the 'classical' tiling is clearly not shift-invariant.

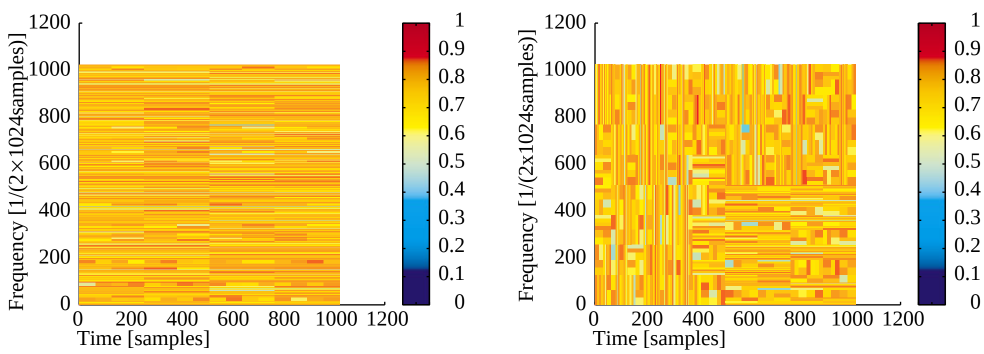

Figure 12. 'Classical' WP representation (left) versus optimal representation (right) of a Gaussian white noise signal (1024 values). The images depict the fact that while the single-tree WP algorithm masks the actual time–frequency symmetry, the absolute optimal representation reveals it.

supported by NSF (grant EAR-9117866), NOAA (grant NA46GP0486) and NASA (grant NAG5-1208). The rainfall data were collected at the Iowa Institute of Hydraulic Research, under the supervision of Konstantine Georgakakos.

Appendix. Generalization of basis optimality results for any concave objective function

The relationship between a concave objective function (e.g. entropy) applied to a basis and the compression performance of that basis is analysed herein. Suppose we have a series of n samples and a normalized basis $\boldsymbol{x}$ thereof, which we order non-increasingly by absolute value, i.e. $|x^{(1)}| \geqslant |x^{(2)}| \geqslant \cdots \geqslant |x^{(n)}|$. Let us denote $\sigma_k = \sum_1^k (x^{(i)})^2$ (usually called 'the sum of the squares of the k most significant basis elements'), and $\sigma_0 = 0$ by convention. Note that the non-negativity of $(x^{(i)})^2$ makes σ_k non-decreasing with k, while the non-increasing ordering of $|x^{(i)}|$ implies that the σ-curves are non-convex, in the sense that $\sigma_k - \sigma_{k-1} \geqslant \sigma_{k+1} - \sigma_k$, $\forall k \in \{1, \ldots, n-1\}$. Also, $\sigma_n = 1$ from the normalization condition.

First, let us consider two bases $\boldsymbol{x} \neq \boldsymbol{y}$, of which $\boldsymbol{y}$ is better than $\boldsymbol{x}$ in compression performance, i.e. $\sigma_k(\boldsymbol{y}) \geqslant \sigma_k(\boldsymbol{x})$, $\forall k \in \{1, \ldots, n\}$. We shall prove that $S(\boldsymbol{y}) \stackrel{\text{Def}}{=} -2\sum_1^n (y^{(i)})^2 \ln |y^{(i)}| < S(\boldsymbol{x})$. To this end, we will devise a finite number of steps that take $|\mathbf{x}|$ into $|\mathbf{y}|$, by decreasing its entropy S at each step. Let $\delta ::= \min_{k \in \{l,\ldots,m-1\}} [\sigma_k(\boldsymbol{y}) - \sigma_k(\boldsymbol{x})]$, such that $\delta > 0$, $\sigma_{l-1}(\boldsymbol{x}) = \sigma_{l-1}(\boldsymbol{y})$ and $\sigma_m(\boldsymbol{x}) = \sigma_m(\boldsymbol{y})$, $n \geqslant m > l > 0$. There has to exist such a subsequence $l, \ldots, m-1$, since $\sigma_0(\boldsymbol{x}) = \sigma_0(\boldsymbol{y}) = 0$, $\sigma_n(\boldsymbol{x}) = \sigma_n(\boldsymbol{y}) = 1$, and initially $\boldsymbol{x} \neq \boldsymbol{y}$. As a first step, let us change $\boldsymbol{x}$ such that $(x^{(l)})^2 ::= (x^{(l)})^2 + \delta$ and $(x^{(m)})^2 ::= (x^{(m)})^2 - \delta$, which translates into $\sigma_k ::= \sigma_k + \delta$, $\forall k \in \{l, \ldots, m-1\}$. Indeed, we find a decrease in entropy: $-[(x^{(l)})^2 + \delta] \ln[(x^{(l)})^2 + \delta] - [(x^{(m)})^2 - \delta] \ln[(x^{(m)})^2 - \delta] < -2(x^{(l)})^2 \ln |x^{(l)}| - 2(x^{(m)})^2 \ln |x^{(m)}|$, since $\delta > 0$, $|x^{(l)}| \geqslant |x^{(m)}|$, and the function $-u \ln u$ is concave from $-(u \ln u)'' = -1/u < 0$, $\forall u > 0$. At this point, for at least one index $k_0 \in \{l, \ldots, m-1\}$ we have $\sigma_{k_0}(\boldsymbol{y}) = \sigma_{k_0}(\boldsymbol{x})$. We can now repeat the above step on subsequences ending, and respectively beginning, at k_0. Non-convexity and monotonicity of $\boldsymbol{\sigma}$ will be respected within the subsequences, since only a vertical translation takes place, and will also be respected around k_0, since $\sigma_{k_0-1}(\boldsymbol{x}) \leqslant \sigma_{k_0-1}(\boldsymbol{y}) \leqslant \sigma_{k_0}(\boldsymbol{y}) = \sigma_{k_0}(\boldsymbol{x}) \leqslant \sigma_{k_0+1}(\boldsymbol{x}) \leqslant \sigma_{k_0+1}(\boldsymbol{y})$. Therefore the existence and non-increasing ordering of $\boldsymbol{x}$ is preserved at each step, making the decrease in entropy happen at the respective next step, until $\max_{k \in \{1,\ldots,n\}} [\sigma_k(\boldsymbol{y}) - \sigma_k(\boldsymbol{x})] = 0 \Longleftrightarrow \boldsymbol{\sigma}(\boldsymbol{y}) = \boldsymbol{\sigma}(\boldsymbol{x})$. Note that $\boldsymbol{\sigma}(\boldsymbol{y}) = \boldsymbol{\sigma}(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{y} = \boldsymbol{x}$ and the algorithm stops (after having gone through a number of steps less than or equal to n).

The above shows that compression performance and entropy introduce the same ordering in a given set of normalized bases, if the compression-performance ordering exists as defined above (i.e. the σ-curves do not cross). As a consequence, the best compression will be achieved by the least-entropy basis. It is important to note two facts:

(1) In the proof we have only used the concavity of $-u \ln u$, therefore *any concave function* of $(x^{(i)})^2$ on $[0, 1]$ introduces the *same* ordering, if the compression-performance ('contrast') ordering exists.

(2) Two orderings can be compared only if they apply to the same set (of bases, in our case). While this may at first sound trivial, let us for instance recall that the single-tree and the double-tree wavepacket algorithms do *not* operate on the same set of bases, since the double tree can access quite a few more bases than the single tree.

It must be noted that the best compression of a signal may not exist at all as defined, in the sense that either of two or more bases will perform best in terms of recovered energy, depending on the *number* of highest basis elements chosen. To illustrate this situation consider the case of the sequence $(2, 3, 1, 9)$, which normalized gives $\boldsymbol{x} = (\frac{2}{\sqrt{95}}, \frac{3}{\sqrt{95}}, \frac{1}{\sqrt{95}}, \frac{9}{\sqrt{95}})$, and its first Haar wavepacket decomposition level is $\boldsymbol{y} =$

$(\frac{5}{\sqrt{190}}, -\frac{1}{\sqrt{190}}, \frac{10}{\sqrt{190}}, -\frac{8}{\sqrt{190}})$. We have $\boldsymbol{\sigma}(\boldsymbol{x}) = (\frac{81}{95}, \frac{90}{95}, \frac{94}{95}, 1)$ and $\boldsymbol{\sigma}(\boldsymbol{y}) = (\frac{50}{95}, \frac{82}{95}, \frac{94.5}{95}, 1)$. Needless to say, reconstruction with one or two basis elements is better when basis $\boldsymbol{x}$ is used (which, by the way, is the single-tree wavepacket, entropy-objective best basis in this case), whereas with three basis elements it is better when basis $\boldsymbol{y}$ is used. Since for such cases it is not clear which the best compression is in the first place, we cannot look for any relation between entropy (or any other objective function) and compression.

References

[1] Cohen L 1995 *Time–Frequency Analysis* (Englewood Cliffs, NJ: Prentice-Hall)

[2] Herley C, Kovačević J, Ramchandran K and Vetterli M 1993 Tilings of the time–frequency plane: construction of arbitrary orthogonal bases and fast tiling algorithms *IEEE Trans. Image Process.* **41** 3341–59

[3] Wickerhauser M V 1991 Lectures on wavelet packet algorithms *Preprint* Department of Mathematics, Washington University

[4] Marr D 1982 *Vision* (San Francisco, CA: Freeman)

[5] Arnéodo A, Argoul F, Bacry E, Elezgaray J, Freysz E, Grasseau G, Muzy J F and Pouligny B 1992 Wavelet transform of fractals *Wavelets and Some of their Applications, Proceedings* ed Y Meyer (Berlin: Springer) pp 286–352)

[6] Coifman R, Meyer Y and Wickerhauser M V 1992 Wavelet analysis and signal processing *Wavelets and Their Applications* ed M B Ruskai *et al* (Boston, MA: Jones and Bartlett) pp 153–78

[7] Shannon C E 1993 *Collected Papers of Claude Elwood Shannon* ed N J A Sloane and A D Wyner (New York: IEEE)

[8] Herley C, Kovačević J, Ramchandran K and Vetterli M 1992 *Arbitrary Orthogonal Tilings of the Time–Frequency Plane (Int. Symp. on Time–Frequency and Time-Scale Analysis 11–14, Victoria, BC)* (New York: IEEE)

[9] Ramchandran K and Vetterli M 1993 Best wavelet packet bases in a rate-distortion sense *IEEE Trans. Image Process.* **41** 160–73