Translation-Invariant Denoising Using the Minimum Description Length Criterion

Israel Cohen*, Shalom Raz and David Malah

Department of Electrical Engineering, Technion — Israel Institute of Technology,
Technion City, Haifa 32000, Israel

June 15, 1998

Abstract

A translation-invariant denoising method, based on the *Minimum Description Length* (MDL) criterion and tree-structured best-basis algorithms is presented. A collection of signal models is generated using an *extended* library of orthonormal wavelet-packet bases, and an additive cost function, approximately representing the MDL principle, is derived. We show that the minimum description length of the noisy observed data is achieved by utilizing the *Shift-Invariant Wavelet Packet Decomposition* (SIWPD) and thresholding the resulting coefficients. This approach is extendable to local trigonometric decompositions, and corresponding procedures to optimize either the library of bases or the filter banks used at each node of the expansion-tree are described. The signal estimator is efficiently combined with a *modified Wigner distribution*, yielding robust time-frequency representations, characterized by high resolution and suppressed interference-terms. The proposed method is compared to alternative existing methods, and its superiority is demonstrated by synthetic and real data examples.

Keywords: Denoising; Signal estimation; Shift-invariant; Wavelet packet; Minimum description length; Best basis; Time-frequency representation; Wigner distribution

*Corresponding author. Tel.: 972 4 879 5033; fax: 972 4 879 5315; e-mail: cisrael@shoshan.technion.ac.il.

1 Introduction

The use of wavelet bases for estimating noisy signals has been the object of considerable recent research. Traditional methods often entail noise removal by low-pass filtering, thus blurring sharp signal features. In contrast, wavelet-based methods show good performance for a wide diversity of signals, including those containing jumps, spikes and other nonsmooth features [17, 11, 12]. The *wavelet shrinkage* method (Donoho and Johnston [19]) is based on transforming the noisy data into a fixed wavelet basis, where soft or hard thresholding is applied to the resulting coefficients. The subsequent synthesis yields the desired signal. It was recognized that such denoising scheme is practically restricted by the extent to which the transform compresses the unknown signal into few significant coefficients [18]. Accordingly, adaptive transforms such as the wavelet packet and local trigonometric decompositions (WPD, LTD) [10], appear to be quite promising [16, 20, 33].

Several approaches and measures to selecting the “best” basis and threshold value, leading to the best signal estimate, have been proposed. In [16, 20], the adapted basis and threshold selection are based on a criterion of minimum mean-squared error. In [3], a complexity-penalized functional is defined using the same threshold, and a subset of basis functions is chosen from a prescribed collection of waveforms. Saito [33] proposed to use an information-theoretic criterion, the *Minimum Description Length* (MDL) principle [32], for the noise removal. He suggested that the MDL criterion provides the best compromise between the estimation fidelity (noise suppression) and the efficiency of representation (signal compression). Unfortunately, the cost function associated with this method is not additive. Thus, he employed the Shannon entropy as the primary cost function for determining the best basis, and the MDL principle merely as a secondary criterion. In [21, 25], the MDL principle is further investigated to derive efficient procedures for selecting the best basis as well as the threshold values. They show that it is possible to define an additive “denoising” criterion so that the conventional WPD remains applicable.

Coifman *et al* [12, 2, 33] observed that denoising with the conventional wavelet transform and WPD may exhibit visual artifacts, such as pseudo-Gibbs phenomena in the neighborhood of discontinuities and artificial symmetries across segmentation points in the frequency domain.

These artifacts are related to the shift-variant representation, and therefore can be reduced by averaging the translation dependence: applying a range of shifts to the noisy data, denoising the shifted versions with the wavelet transform, then unshifting and averaging the denoised data. This procedure, termed *Cycle-Spinning*, generally yields better visual performance on smooth parts of the signal. However, transitory features may be significantly attenuated [35]. Furthermore, the MDL principle and related information-theoretic arguments cannot be applied.

Another approach to attaining shift-invariance is to optimize the time localization of the signal, so that its features are well-aligned with the basis-functions. In the case of WPD, Pesquet *et al.* [28, 29] suggested to adapt the shift of the signal as follows: (*i*) To each node of the expansion tree assign an information-cost by averaging the Shannon entropy over all translations. (*ii*) Determine the best expansion tree using the conventional WPD algorithm of Coifman and Wickerhauser [10]. (*iii*) Compare the entropy of the 2^κ orthonormal representations resulting from 2^κ different shift-options, where κ is the number of nodes in the best expansion tree, and choose that representation (shift-option) which minimizes the entropy. This procedure is sub-optimal compared with the *Shift-Invariant Wavelet Packet Decomposition* (SIWPD) [5, 6], since the expansion tree is determined by the *averaged* entropy. Additionally, the shift-options in step (*iii*) are examined one by one, whereas the SIWPD not only provides a *recursive* selection method for the optimal shift, but also offers an inherent trade-off between the computational complexity and the information cost.

In this paper, we present a translation-invariant denoising method, based on the SIWPD and the MDL criterion. An *extended* library of wavelet-packet bases [6] is employed for generating a collection of competing models, and the MDL principle is applied for approximating the description length of the observed noisy data. We show that minimum description length is attainable by optimizing the expansion-tree associated with the SIWPD. The optimal signal estimate is subsequently obtained by thresholding the resulting coefficients. The proposed method is extendable to other adaptive transforms, *e.g.*, the *Shift-Invariant Adaptive-Polarity Local Trigonometric Decomposition* (SIAP-LTD) [8]. A corresponding procedure to optimize either the library of bases or the filter banks used at each node of the expansion-tree is described as well. The signal estimator is independent of the alignment of the observed signal with respect to the basis functions. Furthermore, the intrinsic advantages of the SIWPD and SIAP-LTD over the conventional WPD and LTD are instrumental in generating a relatively superior estimator.

The proposed algorithm is also useful for estimating the time-frequency distributions of noisy signals. Since the Wigner distribution is very sensitive to noise, it is often necessary to employ some kind of smoothing to reduce the noise effects [4, 27]. However, smoothing suppresses noise at the expense of considerable "smearing" of the signal components. The combination of the above mentioned signal estimator with the recently introduced modified Wigner distribution [9] yields a distribution that is robust to noise and characterized by high resolution, high concentration and suppressed interference-terms.

This paper is organized as follows. In Section 2, we review the SIWPD and demonstrate its shift-invariant properties. In Section 3, we formulate our problem. Specifically, signal estimation is described as a problem of choosing the best model from a collection defined by an extended library of wavelet packet bases. In Section 4, the MDL principle is applied to determine the description length of the data. We show that minimum description length is attainable by optimizing the expansion-tree. In Section 5, we present a corresponding algorithm for the optimal tree design and signal estimation. We also propose an MDL-based estimator for structuring the time-frequency distribution. Examples illustrating the execution and performance of the proposed algorithms are presented in Section 6. The connections between these algorithms and other approaches are discussed in Section 7.

2 The Shift-Invariant Wavelet Packet Decomposition

The SIWPD [6] is an adaptive representation in an *extended* library of wavelet packet bases. The extended library is defined as the collection of all translated versions of the ordinary wavelet packet bases. For a prescribed signal, the SIWPD selects the best basis with respect to an additive information cost functional.

Let $\{\psi_n(t) : n \in \mathbb{Z}_+ \ , t \in \mathbb{R}\}$ be a wavelet packet family [10] generated by

$$\psi_{2n}(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_k \psi_n(2t-k) \tag{1}$$

$$\psi_{2n+1}(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} g_k \psi_n(2t-k) \tag{2}$$

where $g_k = (-1)^k h_{1-k}$, and $\psi_0(t) \equiv \varphi(t)$ is an orthonormal scaling function, satisfying

$$\langle \varphi(t-p), \varphi(t-q)\rangle = \delta_{p,q}, \quad p,q \in \mathbb{Z}. \tag{3}$$

The extended library of wavelet packets is defined as the collection of all the orthonormal bases which are subsets of

$$\left\{B_{\ell,n,m} \ : \ -L \leq \ell \leq 0,\ 0 \leq n, m < 2^{-\ell}\right\}, \tag{4}$$

where $\ell = -L$ denotes the coarsest resolution level, and

$$B_{\ell,n,m} \equiv \left\{\psi_{\ell,n,m,k} = 2^{\ell/2}\psi_n\left(2^\ell(t-m)-k\right) \ : \ 0 \leq k < N2^\ell\right\}. \tag{5}$$

The integer N designates the wavelet packets at the finest resolution level ($\ell = 0$), which are relevant to analyzing the given signal. The extended library is larger than the standard wavelet packet library by a square power, but is still structured into a tree configuration which supports fast search algorithms [5]. The tree is depicted in Fig. 1. Each node in the tree is indexed by the triplet (ℓ, n, m) and represents the subspace

$$U_{\ell,n,m} = \overline{Span}\left\{B_{\ell,n,m}\right\}. \tag{6}$$

Since there are two alternatives for decomposing $U_{\ell,n,m}$ into two orthogonal subspaces:

$$U_{\ell,n,m} = U_{\ell-1,2n,m_c} \oplus U_{\ell-1,2n+1,m_c}, \quad m_c \in \{m, m+2^{-\ell}\}, \tag{7}$$

upon expanding a prescribed node, with minimization of the information cost in mind, we examine and select one of these two alternative decompositions. The branches in the expansion tree are depicted by either fine or heavy lines (Fig. 2), depending on the adaptive selection of m_c.

Let $\mathcal{B}$ and $\mathcal{M}$ represent, respectively, a library of bases and an additive cost function, let $g \in U_{0,0,0}$, and denote by $\mathcal{M}(Bg)$ the information cost of representing g in a basis $B \in \mathcal{B}$.

Definition 1 [10] *The best basis for g in $\mathcal{B}$ with respect to $\mathcal{M}$ is $B \in \mathcal{B}$ for which $\mathcal{M}(Bg)$ is minimal.*

Denote by $A_{\ell,n,m}$ the best basis for g restricted to the subspace $U_{\ell,n,m}$. Then, the SIWPD selects the best basis $A_{0,0,0}$ by the following recursive procedure:

$$A_{\ell,n,m} = \begin{cases} B_{\ell,n,m} \text{ if } \mathcal{M}(B_{\ell,n,m}g) \leq \mathcal{M}(A_{\ell-1,2n,m_c}g) + \mathcal{M}(A_{\ell-1,2n+1,m_c}g), \\ A_{\ell-1,2n,m_c} \oplus A_{\ell-1,2n+1,m_c}, \text{ otherwise,} \end{cases} \tag{8}$$

where the shift indices of the respective children-nodes are obtained by

$$m_c = \begin{cases} m, \text{ if } \sum_{i=0}^{1} \mathcal{M}(A_{\ell-1,2n+i,m}g) \leq \sum_{i=0}^{1} \mathcal{M}(A_{\ell-1,2n+i,m+2^{-\ell}}g) \\ m + 2^{-\ell}, \quad \text{otherwise.} \end{cases} \tag{9}$$

At the coarsest resolution level $\ell = -L$ the subspaces $U_{-L,n,m}$ are not further decomposed, *i.e.*, $A_{-L,n,m} = B_{-L,n,m}$ for $0 \leq n, m < 2^L$.

Compared with the ordinary WPD [10], the SIWPD is determined to be advantageous in the following respects [6]: 1) Shift-invariance; 2) Lower information cost; 3) Improved time-frequency resolution; 4) More stable information cost across a prescribed data set; 5) Controlled computational complexity (at the expense of the information cost down to $O(Nlog_2 N)$). These desirable properties advance signal analysis, compression, identification and classification applications. To illustrate the shift-invariant properties of the SIWPD and its enhanced time-frequency representation compared to the standard WPD, we refer to the expansion of the signal $g(t)$ (Fig. 3) and $g(t - 2^{-6})$. These signals contain $2^7 = 128$ samples, and are identical to within 2 samples time-shift. For definiteness, we choose D_8 to serve as the scaling function (D_8 corresponds to 8-tap Daubechies least asymmetric wavelet filters [13, page 198]) and the Shannon entropy as the cost function, defined by [10] $\mathcal{M}(\{x_i\}) = -\sum_{i:x_i \neq 0} x_i^2 \log x_i^2$. Figs. 4 and 5 display the best-basis expansions under the WPD and the SIWPD algorithms, respectively. The sensitivity of WPD to temporal shifts is obvious, while the best-basis SIWPD representation is indeed shift-invariant and characterized by a lower entropy and improved time-frequency resolution.

3 Problem Formulation

We assume the following model for signal estimation:

$$y(t) = f(t) + z(t) \tag{10}$$

where $y(t)$ represents the noisy observed data, $f(t)$ is the unknown signal to be estimated, and $z(t)$ is a white Gaussian noise (WGN) with zero mean and a presumingly known power spectral density (PSD) σ^2. We assume that $f(t)$ is real-valued and belongs to V_0, where

$$V_0 = \overline{Span}\, \{\psi_0(t-k) \ : \ k \in \mathbb{Z}\} \ , \tag{11}$$

so that Eq. (10) can be projected onto V_0 (this assumption amounts to some weak regularity condition on $f(t)$ [22]). Furthermore, $f(t)$ is assumed to have a compact support, so that there exists a finite integer N such that

$$\langle f \, , \, \psi_{\ell,n,m,k} \rangle = 0 \quad \text{if } k < 0 \text{ or } k \geq N 2^{\ell} \tag{12}$$

where

$$\psi_{\ell,n,m,k}(t) \equiv 2^{\ell/2} \psi_n \left(2^{\ell}(t-m) - k \right) \ , \tag{13}$$

$-log_2 N \leq -L \leq \ell \leq 0$, $0 \leq n, m < 2^{-\ell}$ (N represents the number of wavelet packet coefficients retained at the finest resolution level $\ell = 0$).

To estimate $f(t)$ from the noisy signal $y(t)$, we employ the extended library of wavelet packet bases. Each basis in the library is associated with a tree-set E, that comprises the terminal-nodes indices of a SIWPD tree [6].

Definition 2 *A collection of indices* $E = \{(\ell, n, m) \ : \ -L \leq \ell \leq 0, \quad 0 \leq n, m < 2^{-\ell}\}$ *is called a tree-set if it satisfies*

(i) The segments $I_{\ell,n} = [2^{\ell} n, \ 2^{\ell}(n+1))$ *are a disjoint cover of* $[0, \ 1)$.

(ii) The shift indices of a pair of nodes (ℓ_1, n_1, m_1), $(\ell_2, n_2, m_2) \in E$ *are related by*

$$m_1 \mod 2^{-\hat{\ell}+1} = m_2 \mod 2^{-\hat{\ell}+1} \tag{14}$$

where $\hat{\ell}$ *is the level index of a dyadic interval* $I_{\hat{\ell},\hat{n}}$ *that contains both* I_{ℓ_1,n_1} *and* I_{ℓ_2,n_2}.

By Proposition 1 in [6], $\{B_{\ell,n,m} : (\ell, n, m) \in E\}$ is an orthonormal basis for $U_{0,0,0}$, and the collection of all tree-sets E as specified above generates an extended library of orthonormal wavelet packet bases. Eq. (12) implies that $f(t)$ belongs to $U_{0,0,0} \subset V_0$. Consequently, $f(t)$ can be estimated from

$$\left\{\langle y \, , \, \psi_{\ell,n,m,k} \rangle \; : \; (\ell, n, m) \in E, \, 0 \le k < N2^{\ell}\right\} .$$

Since the bases in the extended library compress signals very well and the tree-set E is adapted to the signal, it is reasonable to assume that $f(t)$ is adequately represented by a small number $K < N$ of orthogonal directions. Accordingly, we consider a signal estimate of the form

$$\hat{f}(t) = \sum_{k=1}^{K} f_k \phi_k(t) \tag{15}$$

where

$$\phi_k \in \{B_{\ell,n,m} : (\ell, n, m) \in E\} . \tag{16}$$

The problem is to find the best tree-set E and the best number of terms K (best model) such that the estimate (15) is optimal according to the MDL principle.

4 The Minimum Description Length Principle

The MDL principle [30, 31, 32] asserts that given a data set and a collection of competing models, the best model is the one that yields the minimal description length of the data. The description length of the data is counted for each model in the collection as the codelength (in bits) of encoding the data using that model, and the codelength needed to specify the model itself. The rationale is that a good model is judged by its ability to "explain" the data, hence the shorter the description length, the better the model.

In order to apply the MDL principle to our problem, we compute the codelength required to encode the data $y(t)$ using the following model

$$y(t) = \sum_{k=1}^{N} y_k \phi_k(t) \,, \tag{17}$$

$$f(t) = \sum_{k=1}^{N} f_k \phi_k(t) \,, \qquad f_k \neq 0 \quad \text{iff } k \in \{k_n\}_{1 \leq n \leq K} \,, \tag{18}$$

$$\{\phi_k \ : \ 1 \leq k \leq N\} = \{B_{\ell,n,m} \ : \ (\ell, n, m) \in E\} \ , \tag{19}$$

$$y_k = f_k + z_k \,, \qquad 1 \leq k \leq N \tag{20}$$

where $y_k = \langle y, \phi_k \rangle$ and $f_k = \langle f, \phi_k \rangle$ are, respectively, expansion coefficients of the observed data and the unknown signal, and $z_k = \langle z, \phi_k \rangle$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ by the orthonormality of the transform. The encoding, and hence the computation of the codelength, is carried out in three steps: (i) encoding the observed data assuming E, K and $\{k_n\}_{1 \leq n \leq K}$ are given; (ii) encoding the number of signal terms K and their locations $\{k_n\}_{1 \leq n \leq K}$ assuming that E is given; and (iii) encoding the tree-set E. Accordingly, the total description length of the data is given by

$$\mathcal{L}(y) = \mathcal{L}\left(y \mid E, K, \{k_n\}_{1 \leq n \leq K}\right) + \mathcal{L}\left(K, \{k_n\}_{1 \leq n \leq K} \mid E\right) + \mathcal{L}(E) \,. \tag{21}$$

We start with the encoding of the observed data assuming E, K and $\{k_n\}_{1 \leq n \leq K}$ are given. It was established by Rissanen [32, pp. 56, 87] that the shortest codelength for encoding the data set $\{y_k\}_{1 \leq k \leq N}$ using the probabilistic model $P(\{y_k\}_{1 \leq k \leq N} \mid \mu)$, where μ is an unknown parameter vector, is asymptotically given by

$$\mathcal{L}(\{y_k\}_{1 \leq k \leq N}) = -\log_2 P(\{y_k\}_{1 \leq k \leq N} \mid \hat{\mu}) + \frac{q}{2} \log_2 N \tag{22}$$

where $\hat{\mu}$ is the maximum likelihood estimator of μ:

$$\hat{\mu} = \arg\max_{\mu} P(\{y_k\}_{1 \leq k \leq N} \mid \mu) \tag{23}$$

and q is the number of free real parameters in the vector μ.

Recalling that the expansion coefficients of the noise $\{z_k\}_{1\le k\le N}$ are i.i.d. $\mathcal{N}(0,\sigma^2)$, it follows from Eq. (20) that the probability of observing the data given all model parameters is,

$$P\left(y \mid \mu\right) = (2\pi\sigma^2)^{-N/2}\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{n=1}^{K}(y_{k_n}-f_{k_n})^2+\sum_{n=K+1}^{N} y_{k_n}^2\right)\right) \tag{24}$$

where

$$\mu = (E, K, \{k_n\}_{1\le n\le K}, \{f_{k_n}\}_{1\le n\le K}) \tag{25}$$

is the parameter vector, and

$$\{k_n\}_{K+1\le n\le N} = \{1,\ldots,N\}\backslash\{k_n\}_{1\le n\le K}\,. \tag{26}$$

Thus, from Eq. (22), the codelength required to encode the observed data, assuming E, K and $\{k_n\}_{1\le n\le K}$ are given, is

$$\begin{aligned}\mathcal{L}\left(y \mid E, K, \{k_n\}_{1\le n\le K}\right) &= -\log_2 P\left(y \mid E, K, \{k_n\}_{1\le n\le K}, \{\hat{f}_{k_n}\}_{1\le n\le K}\right) + \frac{K}{2}\log_2 N \\ &= \frac{1}{2\sigma^2\ln 2}\sum_{n=K+1}^{N} y_{k_n}^2 + \frac{N}{2}\log_2(2\pi\sigma^2) + \frac{K}{2}\log_2 N\end{aligned} \tag{27}$$

where

$$\hat{f}_{k_n} = y_{k_n}\,, \quad 1 \le n \le K \tag{28}$$

are the maximum likelihood estimates of $\{f_{k_n}\}_{1\le n\le K}$.

Next, we encode the number of signal terms K and their locations $\{k_n\}_{1\le n\le K}$ assuming that E is given. The integer K $(1 \le K \le N)$ requires $\log_2 N$ bits (clearly, if the probability density function for K, $P_K(k)$, is known, then $\mathcal{L}(K) = -\sum_{k=1}^{N} P_K(k)\log_2 P_K(k) \le \log_2 N$). The indices $\{k_n\}_{1\le n\le K}$ can be specified by a binary string of length N containing exactly K 1s. Since there are $\binom{N}{K}$ such possible strings, the codelength is given by

$$\mathcal{L}\left(K, \{k_n\}_{1\le n\le K} \mid E\right) = \log_2 N + \log_2\binom{N}{K} = \log_2\frac{N\cdot N!}{K!(N-K)!} \tag{29}$$

By applying Stirling's formula[1] to the factorials we have

$$\mathcal{L}\,(K, \{k_n\}_{1 \le n \le K} \mid E) = Nh(K/N) - \frac{1}{2}\log_2[K(N-K)] - \frac{1}{12 \ln 2}\left(\frac{\theta_1}{K} + \frac{\theta_2}{N-K}\right) + c \qquad (30)$$

where $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function and θ_1, θ_2 and c are constants independent of K $(0 < \theta_1, \theta_2 < 1)$. For $N \gg K$, ignoring constant terms which are independent of K, the codelength can be approximated by

$$\mathcal{L}\,(K, \{k_n\}_{1 \le n \le K} \mid E) \approx K \log_2 N\,. \qquad (31)$$

Since our goal is to obtain the shortest codelength, the optimal number of signal terms K^* and their optimal locations $\{k_n^*\}_{1 \le n \le K}$ are obtained by minimizing the sum of codelengths given by Eqs. (27) and (31):

$$\begin{aligned} \mathcal{L}\,(y \mid E) &= \frac{1}{2\sigma^2 \ln 2} \sum_{n=K+1}^{N} y_{k_n}^2 + \frac{3K}{2}\log_2 N \\ &= \frac{1}{2\sigma^2 \ln 2}\left[\sum_{n=K+1}^{N} y_{k_n}^2 + \sum_{n=1}^{K}(3\sigma^2 \ln N)\right] \end{aligned} \qquad (32)$$

where the constant terms are discarded. Clearly,

$$\sum_{n=1}^{N} \min\left(y_n^2\,,\, 3\sigma^2 \ln N\right) \le \sum_{n=K+1}^{N} y_{k_n}^2 + \sum_{n=1}^{K}(3\sigma^2 \ln N) \qquad (33)$$

for all $1 \le K \le N$ and $\{k_n\}_{1 \le n \le K} \subset \{1, \ldots, N\}$. Equality in (33) holds for the optimal values given by

$$K^* = \#\left\{y_n^2 > 3\sigma^2 \ln N \mid 1 \le n \le N\right\} \qquad (34)$$

and

$$\{k_n^*\}_{1 \le n \le K^*} = \left\{n \mid y_n^2 > 3\sigma^2 \ln N\,,\, 1 \le n \le N\right\}. \qquad (35)$$

[1] $x! = \sqrt{2\pi}\, x^{x+1/2} \exp(-x + \frac{\theta}{12x}) \quad (x > 0,\ 0 < \theta < 1)$

Specifically, given E we compute the expansion coefficients of the observed data, and then K^* is the number of coefficients exceeding the threshold $\sigma\sqrt{3\ln N}$ in absolute value, and $\{k_n^*\}_{1\le n\le K^*}$ are their locations (notice that $K^*=0$ implies $\hat{f}\equiv 0$). Thus the codelength in Eq. (32) reduces to

$$\mathcal{L}\left(y\,|\,E\right)=\frac{1}{2\sigma^2\ln 2}\sum_{n=1}^{N}\min\left(y_n^2\,,\,3\sigma^2\ln N\right)\,. \tag{36}$$

To encode the tree-set E, we associate a 3-ary string with the SIWPD tree as follows: For each node (ℓ,n,m), use 0 if its shift-index m is identical to the shift-index of its child-nodes; use 1 if its child-nodes, $(\ell-1,2n,m_c)$ and $(\ell-1,2n+1,m_c)$, have a different shift-index $(m_c\neq m)$; and use 2 if it is a terminal-node $((\ell,n,m)\in E)$. Now, traverse the tree from node to node, top-down from left to right, starting at the root at the top. The string for the example shown in Fig. 6 is 0210222.

A SIWPD tree includes $|E|$ terminal nodes and $|E|-1$ internal nodes, where $|E|$ is the cardinality of E. Since the tree always ends with a terminal node, the last 2 in the string can be discarded, and thus we need to encode a sequence containing $|E|-1$ 2s and $|E|-1$ symbols from $\{0,1\}$. The description length of such sequence is

$$\mathcal{L}\left(E\right)=\log_2\begin{pmatrix}2|E|-2\\|E|-1\end{pmatrix}+\left(|E|-1\right)+\log_2|E|\,, \tag{37}$$

where the first term is required to specify the locations of 2s in the sequence, the second term to discriminate between 0s and 1s, and the third term to encode the number of terminal terms. Applying Stirling's formula to the factorials, the description length of the tree is given by

$$\mathcal{L}\left(E\right)=3|E|+\log_2\frac{|E|}{\sqrt{|E|-1}}+\frac{\alpha_1-4\alpha_2}{24(|E|-1)\ln 2}+c' \tag{38}$$

where α_1,α_2 and c' are constants independent of E $(0<\alpha_1,\alpha_2<1)$. For $|E|\gg 1$, the codelength can be approximated by

$$\mathcal{L}\left(E\right)\approx 3|E| \tag{39}$$

where the constant terms are ignored. Adding the codelength $\mathcal{L}\left(y\,|\,E\right)$ (Eq. (36)), the total description length of the observed data is given by

$$\mathcal{L}\left(y\right)=\mathcal{L}\left(E\right)+\mathcal{L}\left(y\,|\;E\right)=3\left|E\right|+\frac{1}{2\sigma^2\ln 2}\sum_{n=1}^{N}\min\left(y_n^2\;,\;3\sigma^2\ln N\right)\;. \tag{40}$$

Observe that the dependence of $\mathcal{L}\left(y\right)$ on the tree-set E is introduced through the number of terminal nodes and the values of the expansion coefficients $\{y_n\}_{1\leq n\leq N}$. Since the total energy of the coefficients $\sum_{n=1}^{N}y_n^2=||y||^2$ is independent of E, we want that the relative energy contained in the coefficients exceeding $\sigma\sqrt{3\ln N}$ in magnitude will be as large as possible. At the same time, we want to minimize the complexity of the expansion tree (the number of terminal nodes). In the next section we show that the SIWPD can be utilized for choosing the best E such that $\mathcal{L}(y)$ is minimized.

5 The Optimal Tree Design and Signal Estimation

Let $\mathcal{B}$ represent the extended library of wavelet packet bases. Since each basis B in the library is related to a tree-set E by

$$B=\{B_{\ell,n,m}\;:\;(\ell,n,m)\in E\}\;, \tag{41}$$

the search for the optimal E is equivalent to the search for the optimal basis in $\mathcal{B}$. Denote by $\mathcal{L}(By)$ the description length of y represented on a basis B. Then, by Eq. (40)

$$\mathcal{L}(By)=\sum_{(\ell,n,m)\in E}\mathcal{L}(B_{\ell,n,m}y) \tag{42}$$

where

$$\mathcal{L}(B_{\ell,n,m}y)=3+\frac{1}{2\sigma^2\ln 2}\sum_{k=1}^{2^\ell N}\min\left\{C_{\ell,n,m,k}^2(y)\;,\;3\sigma^2\ln N\right\} \tag{43}$$

is the codelength for the terminal node $(\ell,n,m)\in E$, and

$$B_{\ell,n,m}y=\left\{C_{\ell,n,m,k}(y)=\langle y\;,\;\psi_{\ell,n,m,k}\rangle\;:\;1\leq k\leq 2^\ell N\right\} \tag{44}$$

are the expansion coefficients of the observed data.

Definition 3 *The optimal basis for y in $\mathcal{B}$ with respect to the MDL principle is $B \in \mathcal{B}$ for which $\mathcal{L}(By)$ is minimal.*

The codelength in Eq. (42) is an additive cost function, which directly results from the expressions and approximations derived in the previous section. Accordingly, we can apply the SIWPD on the observed data y, as described in Section 2, in order to find its optimal basis.

The optimal basis $A \equiv A_{0,0,0}$ minimizes the description length of the observed data. Thus, from Eqs. (28), (34) and (35), the optimal estimate of $f(t)$ is obtained by expanding the observed data $y(t)$ on the optimal basis $A = \left\{\hat{\phi}_k\right\}_{1\leq k\leq N}$ and *hard-thresholding* the coefficients by $\tau \equiv \sigma\sqrt{3\ln N}$. Specifically,

$$\hat{f}(t) = \sum_{k=1}^{N} \eta_\tau(y_k)\hat{\phi}_k(t) \tag{45}$$

where $y_k = \left\langle y, \hat{\phi}_k \right\rangle$, and $\eta_\tau(c) \overset{\triangle}{=} c\mathbf{1}_{\{|c|>\tau\}}$ is the *hard-threshold* function.

The signal estimation by the above process is shift-invariant, since the optimal basis expansion obtained by the SIWPD is shift-invariant. Accordingly, if the observed data $y(t)$ is translated in time by $q \in \mathbb{Z}$, then the signal estimate $\hat{f}(t)$ is also translated by q. Observe that the restriction of the translations to integers stems from the fact that the initial (finest) resolution level of representing the observed signal is $\ell = 0$, as the unknown signal $f(t)$ is assumed to be in V_0. If we use a finer resolution level $J > 0$ for the initial discrete representation, the shift-invariance is satisfied for finer translations of the form $2^{-J}q$, where $q \in \mathbb{Z}$. However, the resolution levels $0 < \ell \leq J$ add no information to estimating the signal, and consequently the execution of SIWPD over the resolution levels $\ell > 0$ merely increases the computational complexity without improving the performance of the estimator.

The following steps summarize the execution of translation-invariant denoising using the MDL criterion:

Step 0 *Choose an extended library of wavelet packet bases $\mathcal{B}$ (i.e, specify a mother wavelet for the SWP library) and specify the maximum depth of decomposition L ($L \leq \log_2 N$).*

Step 1 *Expand the data y into the library $\mathcal{B}$. i.e., obtain the coefficients $B_{\ell,n,m}y = \{C_{\ell,n,m,k}(y)\}_{1 \leq k \leq 2^\ell N}$ for $-L \leq \ell \leq 0$, $0 \leq n, m < 2^{-\ell}$.*

Step 2 *Use Eq. (43) to determine $\mathcal{L}(B_{\ell,n,m}y)$ for $-L \leq \ell \leq 0$, $0 \leq n, m < 2^{-\ell}$, and set $A_{-L,n,m} = B_{-L,n,m}$ for $0 \leq n, m < 2^L$.*

Step 3 *Determine the optimal basis $A \equiv A_{0,0,0}$ and the minimum description length $\mathcal{L}(Ay)$ using Eqs. (8)-(9), where $\mathcal{M}(\cdot) \equiv \mathcal{L}(\cdot)$.*

Step 4 *Threshold the expansion coefficients in the selected basis by $\tau = \sigma\sqrt{3 \ln N}$ and reconstruct the signal estimate, as expressed by (45).*

The computational complexity of executing an optimal SIWPD best-basis expansion is $O(N2^{L+1})$. Yet, as demonstrated in [6], one may resort to a *sub-optimal* SIWPD procedure entailing a reduced complexity, and higher description length (*i.e.*, information cost) while still retaining the desirable shift-invariance property. In that case, the depth of a subtree, used at a given parent-node to determine its shift index, is restricted to d resolution levels ($1 \leq d \leq L$), and the computational complexity reduces to $O[2^d(L - d + 2)N]$. In the extreme case $d = 1$, the complexity, $O(NL)$, is similar to that associated with the conventional WPD. The larger d and L, the larger the complexity, however, the determined optimal basis generally yields a shorter description length.

Similar to the algorithm described in [33], our algorithm can also be extended to find the optimal basis in more than one library. Given a collection of libraries $\{\mathcal{B}_i\}_{1 \leq i \leq P}$ including a few extended libraries of wavelet packet and local trigonometric bases, we can find the optimal basis that minimizes the description length as follows: For each library $\mathcal{B}_i$ ($1 \leq i \leq P$), find the optimal basis $A_i \in \mathcal{B}_i$ and the description length $\mathcal{L}(A_i y)$ as described above. Then, choose the optimal basis A such that $\mathcal{L}(Ay) = \min\{\mathcal{L}(A_i y) \,:\, 1 \leq i \leq P\}$. In the case of an extended library of local trigonometric bases [6], the codelength associated with a terminal node is also approximated by Eq. (43). Each node in a SIAP-LTD tree has only two expansion alternatives, for it is either decomposed or selected as a terminal node (in contrast to the SIWPD tree, where each node has three expansion alternatives). However, another bit is required for each terminal node to specify its polarity [6]. Therefore, the description lengths of SIAP-LTD and SIWPD trees are approximately the same.

Finding the optimal basis $A = \left\{\hat{\phi}_k\right\}_{1\le k\le N}$, the signal estimate is once again obtained by Eq. (45). Alternatively, the decomposition filters can be adapted to the statistics of the signal in each node [25]. Joint adaptation of filter banks and tree structures has been utilized in image coding applications [15, 26], and a fast algorithm for maximizing energy compaction was introduced in [24]. In our case, to compute the description length of the observed data, the codelength of an internal node should include the specification of the filters applied to expand it. Since the number of internal nodes is relative to the number of terminal nodes (there are $|E| - 1$ internal nodes and $|E|$ terminal nodes), the MDL can be obtained by adding to $\mathcal{L}(B_{\ell,n,m}y)$ (expression (43)) the codelength required to specify the filter banks. Specifically, the codelength of a terminal node is given by

$$\mathcal{L}(B_{\ell,n,m}y) = \log_2 M + 3 + \frac{1}{2\sigma^2 \ln 2} \sum_{k=1}^{2^\ell N} \min\left\{C^2_{\ell,n,m,k}(y)\ ,\ 3\sigma^2 \ln N\right\}\ , \tag{46}$$

where M is the number of different decomposition filters being examined at each *internal* node.

The proposed algorithm for signal estimation is also useful for estimating the time-frequency distributions of noisy signals. While the conventional Wigner distribution (WD) is very sensitive to noise and smoothing is usually applied to reduce noise at the expense of considerable smearing of the signal components [4, 27], the above signal estimate, combined with the recently introduced modified Wigner distribution (MWD) [9], yields robust time-frequency representations. Denote by W_ϕ the auto WD of ϕ, and by W_{ϕ_1,ϕ_2} the cross WD of ϕ_1 and ϕ_2:

$$W_\phi(t,\omega) \;=\; \int \phi(t+\tau/2)\phi^*(t-\tau/2)e^{-j\omega\tau}\, d\tau\ , \tag{47}$$

$$W_{\phi_1,\phi_2}(t,\omega) \;=\; \int \phi_1(t+\tau/2)\phi_2^*(t-\tau/2)e^{-j\omega\tau}\,d\tau\,. \tag{48}$$

Then, from [9] and Eq. (45), the MWD estimate of y is given by

$$\hat{T}_y(t,\omega) = \sum_{k\in\Lambda} |y_k|^2 W_{\hat{\phi}_k}(t,\omega) + 2 \sum_{\{k,k'\}\in\Gamma} \mathrm{Re}\{y_k y_{k'}^* W_{\hat{\phi}_k,\hat{\phi}_{k'}}(t,\omega)\} \tag{49}$$

where

$$\Lambda = \left\{k \;:\; |y_k| > \sigma\sqrt{3\ln N},\; 1 \le k \le N\right\}, \tag{50}$$

$$\Gamma = \left\{\{k,k'\} \;:\; k,k' \in \Lambda,\; 0 < d(\hat{\phi}_k,\hat{\phi}_{k'}) \le D\right\}. \tag{51}$$

Specifically, the set Λ contains the indices of basis-functions whose coefficients are larger than $\sigma\sqrt{3\ln N}$ in magnitude, and Γ restricts the cross terms to *neighboring* pairs of basis-functions, *i.e.*, basis-functions whose time-frequency distance is smaller than a certain distance-threshold D. The distance measure in the time-frequency plane is defined by

$$d(\hat{\phi}_k,\hat{\phi}_{k'} = \left[\frac{(\bar{t}_k-\bar{t}_{k'})^2}{\Delta t_k \Delta t_{k'}} + \frac{(\bar{\omega}_k-\bar{\omega}_{k'})^2}{\Delta\omega_k\Delta\omega_{k'}}\right]^{1/2} \tag{52}$$

where $(\bar{t}_k,\bar{\omega}_k)$ is the position of the cell associated with $\hat{\phi}_k$; Δt_k and $\Delta\omega_k$ are, respectively, the widths (uncertainties) in time and frequency. Similar notations apply to $\hat{\phi}_{k'}$. The distance threshold is adjusted to balance the cross-term interference, the useful properties of the distribution, and the computational complexity [9]. In the next section we show by examples that the above estimate of the time-frequency distribution is robust to noise and possesses the all useful properties of the modified Wigner distribution, namely high energy concentration, well delineated components, low interference-terms, *etc.*

6 Examples

In this section, we give two examples for demonstrating the execution and performance of the proposed denoising method.

Example 1 *Synthetic signal.*

We created a synthetic signal $f_1(t)$ by a linear superposition of a few wavelet packets, generated by the C_{12} scaling function (C_{12} corresponds to $12-$tap coiflet filters [13, page 261] [14]). The signal contains $N = 2^7$ samples and is depicted in Fig. 7(a). Its SIWPD is illustrated in Fig. 7(b), where the Shannon entropy is used as the cost function. The noisy observation $y_1(t)$ (Fig. 7(c)) was created by adding WGN to $f_1(t)$ with signal-to-noise ratio SNR= 7dB. The optimal SIWPD of $y_1(t)$ using the MDL criterion is shown in Fig. 7(d). Notice the remarkable resemblance between the optimal representation of the noisy signal using the MDL principle and the ordinary SIWPD of the original signal using the Shannon entropy. This resemblance stems from fact that according to the MDL principle, the relative energy, contained in the coefficients exceeding $\sigma\sqrt{3\ln N}$ in magnitude, should be as large as possible (refer to Eq. (40)). While by the Shannon entropy, the expansion coefficients in the best-basis should decrease as rapidly as possible, when rearranged in a decreasing magnitude order. Therefore, the Shannon entropy applied to the original signal and the MDL criterion applied to the noisy signal generally produce similar SIWPD, as long as the threshold level (noise) is lower than the expansion coefficients of the original signal in the best-basis.

Pursuing the estimation procedure with the MDL criterion, the expansion coefficients of $y_1(t)$ in the optimal basis are thresholded by $\sigma\sqrt{3\ln N}$ and transformed back into the signal domain. Figs. 7(e) and (f) show, respectively, the retained coefficients and the signal estimate $\hat{f}_1(t)$. Compared to the noisy measurement $y_1(t)$, the signal estimate is enhanced to SNR= 19dB.

Fig. 8 illustrates the usefulness of our algorithm for estimating the time-frequency distribution of the noisy data. While the WD of the original signal is corrupted by interference terms and even worsens by the noise (Figs. 8(a) and (b)), the Smoothed pseudo Wigner distributions are more readable and less sensitive to noise (Figs. 8(c) and (d)). However, the energy concentration of the signal components is poor. The estimate of the MWD, given by Eq. (49), is not only robust to noise (compare Figs. 8(e) and (f)), but also characterized by high resolution, high concentration and suppressed interference-terms.

Example 2 *Evolution of electromagnetic pulse in a relativistic magnetron.*

Fig. 9(a) shows a noisy measurement of an electromagnetic pulse ($\approx$ 100 nanoseconds long) generated by high power ($\approx$ 100 MegaWatts) relativistic magnetron. The measurement involves heterodyning at 2.6GHz, filtering at 500kHz and sampling at 1GHz [34]. The Wigner distribution, depicted in Fig. 9(b), is clearly ineffective as a time-frequency analysis tool, for its high noise sensitivity. Yet, the estimates of the signal and the MWD, as shown in Figs. 9(c) and (d), are potentially valuable when analyzing the measurements and studying the non-stationary phenomena, such as mode build-up and competition and pulse shortening [1], which are common in such high power microwave tubes.

In this example, we employed the SIAP-LTD [8], since it yielded a shorter description length than the SIWPD (probably because the energy of the pulse is concentrated in the cavity-modes of the magnetron, and local trigonometric bases are more appropriate for describing oscillations). The residual between the noisy measurement and the signal estimate is depicted in Fig. 9(e). To ascertain that this residual is actually the noise component, we compare the estimate of the MWD with the smoothed pseudo Wigner distribution of the noisy measurement (Fig. 9(f)). Since these two distributions are similar, in view of the fact that smoothing in the Wigner domain reduces the noise at the expense of smearing the signal components, it is reasonable to assume that the signal estimate contains all the signal components and the residual is mostly noise.

7 Relation to Other Work

Our algorithm has a close relationship with the "simultaneous noise suppression and signal compression" algorithm developed by Saito [33]. For a given collection of orthonormal bases $\{B_p\}_{1\leq p\leq P}$ consisting of standard wavelet-packet and local trigonometric bases, his algorithm first selects the optimal basis $A \equiv B_{p^*}$ and the optimal number of retained coefficients $K^* < N$ by the MDL principle:

$$\{p^*, K^*\} = \arg \min_{\substack{1 \le p \le P \\ 0 \le K < N}} \left\{ \mathcal{L}(B_p y) = \frac{3}{2} K \log N + \frac{N}{2} \log\left(\sum\nolimits_{k=K+1}^{N} C_{p,k}^2(y)\right) \right\} \tag{53}$$

where $\{C_{p,k}(y) \equiv \langle y, \phi_{p,k} \rangle\}_{1 \le k \le N}$ are the expansion coefficients of y represented in the basis $B_p = \{\phi_{p,k}(t)\}_{1 \le k \le N}$, sorted in order of decreasing magnitude. Then, the signal estimate is reconstructed from the K^* largest expansion coefficients in the optimal basis:

$$\hat{f}(t) = \sum_{k=1}^{K^*} C_{p^*,k}(y) \phi_{p^*,k}(t) \tag{54}$$

(compare Eqs. (53) and (54) with (34) and (45)). To maintain a manageable computational complexity, when considering *libraries* of bases only *one* basis out of each library is being examined, by taking that basis which minimizes the Shannon entropy of the observed data. The main differences between our algorithm and that of Saito are:

- Our method selects the optimal basis by the MDL principle whereas his method first minimizes the Shannon entropy to determine the "best-basis" in each library and only then applies the MDL principle to select the optimal basis among the "best-bases".

- His method ignores the codelength required to specify the best-basis in its library, and thus complex expansion trees are not penalized. On the other hand, our method imposes a significant penalty (up to $3 \cdot 2^L$ bits) for complex trees.

- Our method assumes that the PSD of the noise (σ^2) is known whereas his method estimates it from the $N - K$ smallest coefficients by $\frac{1}{N} \sum_{k=K+1}^{N} C_{p,k}^2(y)$ (maximum-likelihood estimate). In our algorithm we can use different measurements or more advanced methods to estimate the noise, whereas the above estimate of σ^2 heavily relies on the assumption that $f(t)$ is orthogonal to $\{\phi_{p^*,k}(t)\}_{K^*+1 \le k \le N}$.

- Our method translates the MDL criterion into an additive information cost function and thus best-basis search algorithms are applicable, whereas his method computes the description length in each basis one at a time.

Figs. 10–12 demonstrate the comparison between our algorithm and that of Saito, using the synthetic signal analyzed in Example 1. Suppose that the library of bases includes the wavelet packet bases generated by the C_{12} scaling function (recall that the synthetic signal $f_1(t)$ was formed using this library), then according to Saito, the best basis is obtained by a conventional WPD with the Shannon entropy employed as the cost function. The resultant expansion-tree and coefficients of the noisy observation $y_1(t)$ are illustrated in Figs. 10(a) and (b), respectively. Since the compression of the signal by the WPD is insufficient, some of the coefficients containing signal energy are regarded as noise and set to zero. The retained coefficients are shown in Fig. 10(c). The signal estimate, reconstructed from these coefficients, is depicted in Fig. 10(d). Observe that the SNR for the signal estimate got worse than for the noisy measurement (1.1dB< 7dB).

The WPD is a special case of the SIWPD [6]. Therefore, the SIWPD yields sparser representations and better estimates than the WPD, even using the Saito method (compare Figs. 11 and 10). Still, the selection of the best-basis by the Shannon entropy criterion, as discussed above, is not optimal with regard to the MDL principle. The results obtained using our method are depicted in Fig. 12. The expansion of the signal estimate by the MDL principle (Fig. 12(c)) is similar to the expansion of the original signal (Fig. 7(b)). The SNR for the signal estimate is significantly higher than for the noisy measurement (19dB> 7dB).

Our algorithm is also intimately connected to the denoising algorithm of Krim and Pesquet [21]. Their algorithm first applies the WPD to the observed data using the information cost

$$\mathcal{M}(\{y_n\}) = \sum_n \min\left(y_n^2\,,\, 2\sigma^2\log_2 N\right)\,, \tag{55}$$

and then reconstructs the signal estimate from the coefficients that are larger than $\sigma\sqrt{2\log_2 N}$ in magnitude. Their method, however, disregards the description length of the expansion tree (compare Eqs. (55) and (40)). Furthermore, while our method attains shift-invariance by utilizing the SIWPD and SIAP-LTD, their method, restricted by the WPD, admits of signal estimates and performances which are significantly influenced by the alignment of the observation with respect to the basis functions.

Donoho and Johnstone [16] used a different approach to select from a library of bases the "ideal basis" for the signal estimator. Rather than the MDL principle, their criterion was the mean-squared error. They showed that from this point of view, the best-basis for denoising is one minimizing

$$\mathcal{M}(\{y_n\}) = \sum_n \min\left(y_n^2\,,\,\zeta^2\right)\,, \tag{56}$$

where $\zeta = \nu\sigma(1+\sqrt{2\ln M_N})$, M_N is the number of distinct basis-functions contained in the library (for WPD, $M_N = N\log_2 N$) and $\nu > 8$. The signal is then reconstructed in the best-basis from the coefficients which are larger than ζ in magnitude. The threshold ζ is larger than $\tau = \sigma\sqrt{3\ln N}$, obtained by the MDL principle (see Eq. (45)), by at least a factor of $8\sqrt{2/3}$. Thus, the criterion (56) imposes a larger penalty on nonzero coefficients, but nothing for the complexity of the expansion-tree (compare with Eq. (40)).

The methods mentioned above try to recover the signal from a few basis-functions that belong to one of the bases in a library. Alternatively, one could gather all the basis-functions which comprise the library into a *dictionary* of functions, and then search for the "best" reconstruction (not necessarily orthogonal) of the signal estimate according to a specified criterion. Let $\mathcal{D}$ denote an overcomplete dictionary of waveforms, and let

$$\hat{f}(t) = \sum_{k=1}^{N} \hat{f}_k \phi_k(t)\,, \qquad \{\phi_k\}_{1\le k\le N} \subset \mathcal{D} \tag{57}$$

be the signal estimate model. Chen and Donoho [3] proposed to choose the optimal set of elements $\{\phi_k\}_{1\le k\le N}$ and optimal set of coefficients $\{\hat{f}_k\}_{1\le k\le N}$ by solving the penalized problem

$$\min_{\hat{f}} \left\{ \frac{1}{2}\left\|y-\hat{f}\right\|_2^2 + \sigma\xi\cdot\sum_{k=1}^{N}\left|\hat{f}_k\right| \right\} \tag{58}$$

where $\xi = \sqrt{2\ln M_N}$, and M_N is the cardinality of the dictionary. They showed that the solution to this problem can be obtained by linear programming, and compared it by examples to: (*i*) the Donoho-Johnstone estimator described above; (*ii*) the Method-of-Frames denoising (MOFDN), which refers to the solution of

$$\min_{\hat{f}} \left\{ \left\| y - \hat{f} \right\|_2^2 + \xi \cdot \sum_{k=1}^{N} \left| \hat{f}_k \right|^2 \right\} ; \tag{59}$$

and (*iii*) the Matching-Pursuit denoising (MPDN), which runs Matching-Pursuit [23] until the coefficient associated with the selected waveform gets below the threshold ξ. The solution to (58), which was named *Basis-Pursuit* denoising (BPDN), generally results in fewer significant coefficients than the MOFDN, more stable than the MPDN, and outperforms the Donoho-Johnstone estimator when the true signal has a moderate number of nonorthogonal components. However, the BPDN is computationally much more expensive than the other methods.

It is interesting to recognize that part of the criterion in our method, which is based on the MDL principle, is similar to expressions (58) and (59). Inserting Eqs. (18) and (28) into (32), we have that $\mathcal{L}(y \mid E)$, the description length of the noisy data given the expansion-tree, can be written as

$$\mathcal{L}(y \mid E) = \frac{1}{2\sigma^2 \ln 2} \left\{ \left\| y - \hat{f} \right\|_2^2 + \sigma^2 (3 \ln N) \cdot \sum_{n=1}^{K} \left| \hat{f}_{k_n} \right|^0 \right\} . \tag{60}$$

Here, the penalty term includes an ℓ^0 norm of the coefficients, whereas BPDN and MOFDN use ℓ^1 and ℓ^2 norms, respectively. Considering again the estimation problem described in Example 1, Fig. 13 shows the signal estimates of the synthetic signal obtained by the Donoho-Johnstone method, MOFDN, BPDN and MPDN. The dictionary of basis-elements employed in these algorithms is derived from the WPD with the C_{12} scaling function. Compared to the signal estimate in our method (Fig. 7(f)), the above estimates have very low signal-to-noise ratios (Table 1). The deficient recovery of the original signal results from the restricted compression capability of the WPD-dictionary. While the SIWPD optimizes the representation of the signal by incorporating translations of wavelet-packets into the dictionary, the WPD-dictionary is inadequate for signal components that are not aligned with the basis elements. Thus, combing the extended libraries of orthonormal bases with the fast best-basis search algorithms (*e.g.*, the SIWPD and SIAP-LTD), the proposed method facilitates shift-invariant estimators at a manageable computational complexity, which are based on the MDL criterion.

8 Summary

Described herein is a translation-invariant denoising method, which uses the MDL criterion and tree-structured best-basis algorithms. We have defined a collection of signal models based on an extended library of orthonormal bases, and applied the MDL principle to derive a suitable additive cost function. The description length of the noisy observed data was then minimized by utilizing the SIWPD, thus optimizing the expansion-tree associated with the best-basis algorithm, and thresholding the resulting coefficients. Furthermore, the signal estimator was combined with a newly defined modified Wigner distribution, whose time-frequency robustness was amply illustrated. The proposed method was compared to alternative existing methods, and its superiority was demonstrated by synthetic and real data examples.

References

[1] J. Benford and J. Swegle, *High Power Microwaves*, Artech House, Norwood, 1992.

[2] J. Berger, R. R. Coifman and M. Goldberg, "Removing noise from music using local trigonometric bases and wavelet packets", J. Audio Eng. Soc., Vol. 42, Dec. 1994, pp. 808–818.

[3] S. Chen and D. L. Donoho, "Atomic decomposition by basis pursuit", Technical Report, Dept.of Statistics, Stanford Univ., Feb. 1996.

[4] L. Cohen, "Time-frequency distributions — a review", Proc. IEEE, Vol. 77, No. 7, July 1989, pp. 941–981.

[5] I. Cohen, S. Raz and D. Malah, "Shift invariant wavelet packet bases", Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 1081–1084.

[6] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant wavelet packet decomposition and representation", Signal Processing, Vol. 57, No. 3, Mar. 1997, pp. 251–270.

[7] I. Cohen, S. Raz, D. Malah and I. Schnitzer, "Best-basis algorithm for orthonormal shift-invariant trigonometric decomposition", Proc. of the 7th IEEE Digital Signal Processing Workshop, DSPWS'96, Loen, Norway, 1–4 Sep. 1996, 1–4 Sep. 1996, pp. 401–404.

[8] I. Cohen, S. Raz and D. Malah, "Orthonormal shift-invariant adaptive local trigonometric decomposition", Signal Processing, Vol. 57, No. 1, Feb. 1997, pp. 43–64.

[9] I. Cohen, S. Raz and D. Malah, "Eliminating interference terms in the Wigner distribution using extended libraries of bases", Proc. of the 22th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, Germany, 20–24 Apr. 1997, pp. 2133–2136.

[10] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Trans. Inform. Theory, Vol. 38, No. 2, Mar. 1992, pp. 713–718.

[11] R. R. Coifman and F. Majid, "Adapted waveform analysis and denoising", in: Y. Meyer and S. Roques, eds., *Progress in Wavelet Analysis and Applications*, Editions Frontieres, France, 1993, pp. 63–76.

[12] R. R. Coifman and D. L. Donoho, "Translation-invariant de-noising", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995, pp. 125–150.

[13] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM Press,Philadelphia, Pennsylvania, 1992

[14] I. Daubechies, "Orthonormal bases of compactly supported wavelets, II. Variations on a theme", SIAM J. Math. Anal., Vol. 24, No. 2, 1993, pp. 499–519.

[15] P. Delsarte, B. Macq and D. T.M. Slock, "Signal adapted multiresolution transform for image coding", IEEE Trans. Inform. Theory, Vol. 38, No. 2, 1992, pp. 897–904.

[16] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthonormal basis chosen from a library of bases", Comptes Rendus Acad. Sci., Ser. I, Vol. 319, 1994, pp. 1317–1322.

[17] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage", Biometrica, Vol. 81, 1994, pp. 425–455.

[18] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation", Applied and Computational Harmonic Analysis, Vol. 1, 1994, pp. 100–115.

[19] D. L. Donoho, "De-noising by soft thresholding", IEEE Trans. Inform. Theory, Vol. 41, May 1995, pp. 613–627.

[20] H. Krim, S. Mallat, D. Donoho and A. S. Willsky, "Best basis algorithm for signal enhancement", Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 1561–1564.

[21] H. Krim, and J.-C. Pesquet, "On the statistics of best bases criteria", in: A. Antoniadis and G. Oppenheim, ed., *Wavelet and Statistics*, Lecture Notes in Statistics, Springer-Verlag, 1995, pp. 193–207.

[22] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet decomposition", IEEE Trans. PAMI, Vol. 11, No. 7, July 1989, pp. 674–693.

[23] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries", IEEE Trans. on Signal Processing, Vol. 41, No. 12, Dec. 1993, pp. 3397–3415.

[24] P. Moulin, “A new look at signal-adapted QMF bank design”, Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 1312–1315.

[25] P. Moulin, “Signal estimation using adapted tree-structured bases and the MDL principle”, Proc. of the 3rd IEEE-SP Int. Symposium on Time-Frequency and Time-Scale Analysis, Paris, France, 18-21 June 1996, pp. 141–143.

[26] P. Moulin, K. Ramchandran and V. Pavlovic, “Transform image coding based on joint adaptation of filter banks and tree structures”, Proc. Int. Conf. on Image Processing, ICIP’96, Lausanne, Switzerland, Sep. 1996.

[27] A. H. Nuttall, “Wigner distribution function: Relation to short-term spectral estimation, smoothing, and performance in noise”, Naval Underwater Systems Center, Technical Report, No. 8225, 1988.

[28] J.-C. Pesquet, H. Krim, H. Carfantan and J. G. Proakis, “Estimation of noisy signals using time-invariant wavelet packets”, Proc. of Asilomar Conference, Monterey, CA, USA, Vol. 1, Nov. 1993, pp. 31–34.

[29] J.-C. Pesquet, H. Krim and H. Carfantan, “Time-invariant orthonormal wavelet representations”, IEEE Trans. on Signal Processing, Vol. 44, No. 8, Aug. 1996, pp. 1964–1996.

[30] J. Rissanen, “Modeling by shortest data description”, Automatica, Vol. 14, 1978, pp. 465–471.

[31] J. Rissanen, “Universal coding, information, prediction, and estimation”, IEEE Trans. Inform. Theory, Vol. 30, No. 4, July 1984, pp. 629–636.

[32] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, World Scientific, Singapore, 1989.

[33] N. Saito, “Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion”, in: E. Foufoula and P. Kumar, eds., *Wavelets in Geophysics*, Academic Press, 1994, pp. 299–324.

[34] I. Schnitzer, A. Rosenberg, C. Leibovitch, M. Botton, I. Cohen and J. Leopold, “Evolution of spectral power density in grounded cathode relativistic magnetron”, Proc. SPIE, Intense Microwave Pulses IV, Vol. 2843, Aug. 1996.

[35] N. A. Whitmal, J. C. Rutledge and J. Cohen, “Wavelet-based noise reduction”, Proc. of the 20th IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-95, Detroit, Michigan, 8–12 May 1995, pp. 3003–3006.

Table Captions

Table 1: Signal-to-noise ratios for the signal estimates of the synthetic signal using the library of wavelet packets (12-tap coiflet filters) and various denoising methods. The SNR obtained by the proposed *MDL-based Translation-Invariant Denoising* method is significantly higher than those obtained with alternative methods.

Figure Captions

Fig. 1: The extended set of wavelet packets organized in a binary tree structure. Each node in the tree is indexed by the triplet (ℓ, n, m) and represents the subspace $U_{\ell,n,m}$.

Fig. 2: Alternative decompositions of a parent-node (ℓ, n, m) in a SIWPD tree. The branches to the children-nodes $(\ell - 1, 2n, m_c)$ and $(\ell - 1, 2n, m_c)$ are depicted by fine lines if $m_c = m$, and by heavy lines if $m_c = m + 2^{-\ell}$.

Fig. 3: Test signal $g(t)$.

Fig. 4: Effect of a temporal shift on the time-frequency representation using the WPD with 8-tap Daubechies least asymmetric wavelet filters: (a) The best expansion tree of $g(t)$. (b) $g(t)$ in its best basis; Entropy= 2.84. (c) The best expansion tree of $g(t - 2^{-6})$. (d) $g(t - 2^{-6})$ in its best basis; Entropy= 2.59.

Fig. 5: Time-frequency representation using the SIWPD with 8-tap Daubechies least asymmetric wavelet filters: (a) The best expansion tree of $g(t)$. (b) $g(t)$ in its best basis; Entropy= 1.92. (c) The best expansion tree of $g(t - 2^{-6})$. (d) $g(t - 2^{-6})$ in its best basis; Entropy= 1.92. Fine and heavy lines in the expansion tree designate alternative node decompositions. Compared with the WPD (Fig. 4), beneficial properties are shift-invariance and lower information cost.

Fig. 6: Exemplifying the description of SIWPD trees by 3-ary strings. Terminal nodes are represented by 2s, and internal nodes by either 0s or 1s, depending on their expansion mode. In the present example, the string is 0210222.

Fig. 7: Signal estimation by SIWPD and MDL principle: (a) Synthetic signal $f_1(t)$. (b) SIWPD of $f_1(t)$ using the Shannon entropy. (c) Noisy measurement $y_1(t)$; SNR= 7dB. (d) SIWPD of $y_1(t)$ using the MDL principle. (e) The expansion coefficients of $y_1(t)$ after hard-thresholding. (f) The signal estimate $\hat{f}_1(t)$; SNR= 19dB.

Fig. 8: Contour plots of time-frequency distributions: (a) Wigner distribution for the original signal $f_1(t)$. (b) Wigner distribution for the noisy measurement $y_1(t)$. (c) Smoothed pseudo Wigner distribution for $f_1(t)$. (d) Smoothed pseudo Wigner distribution for $y_1(t)$.

(e) The modified Wigner distribution for $f_1(t)$. (f) The estimate of the modified Wigner distribution for $y_1(t)$ by the MDL principle.

Fig. 9: Electromagnetic pulse in a relativistic magnetron (heterodyne detection; local oscillator= 2.6GHz): (a) Noisy measurement $y_2(t)$. (b) Wigner distribution for $y_2(t)$. (c) The signal estimate $\hat{f}_2(t)$ by the MDL principle. (d) The estimate of the modified Wigner distribution for $y_2(t)$. (e) Residual between $y_2(t)$ and $\hat{f}_2(t)$. (f) Smoothed pseudo Wigner distribution for $y_2(t)$.

Fig. 10: Signal estimation by the Saito method using the WPD: (a) The best expansion tree of $y_1(t)$ (the signal is depicted in Fig. 7(c)). (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 1.1dB.

Fig. 11: Signal estimation by the Saito method using the SIWPD: (a) The best expansion tree of $y_1(t)$. (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 12.8dB.

Fig. 12: Signal estimation by the proposed method: (a) The optimal expansion tree of $y_1(t)$. (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 19dB.

Fig. 13: Signal estimates of the synthetic signal using the library of wavelet packets (12-tap coiflet filters): (a) The Donoho-Johnstone method; SNR= 6.4dB. (b) The Method-of-Frames denoising (MOFDN); SNR= 7.1dB. (c) The Basis-Pursuit denoising (BPDN); SNR= 4.3dB. (d) The Matching-Pursuit denoising (MPDN); SNR= 7.5dB.

Denoising Method	**SNR (dB)**
Saito + WPD	1.1
Basis-Pursuit	4.3
Donoho-Johnstone	6.4
Method-of-Frames	7.1
Matching-Pursuit	7.5
Saito + SIWPD	12.8
The proposed method	19.1

Table 1: Signal-to-noise ratios for the signal estimates of the synthetic signal using the library of wavelet packets (12-tap coiflet filters) and various denoising methods. The SNR obtained by the proposed *MDL-based Translation-Invariant Denoising* method is significantly higher than those obtained with alternative methods.

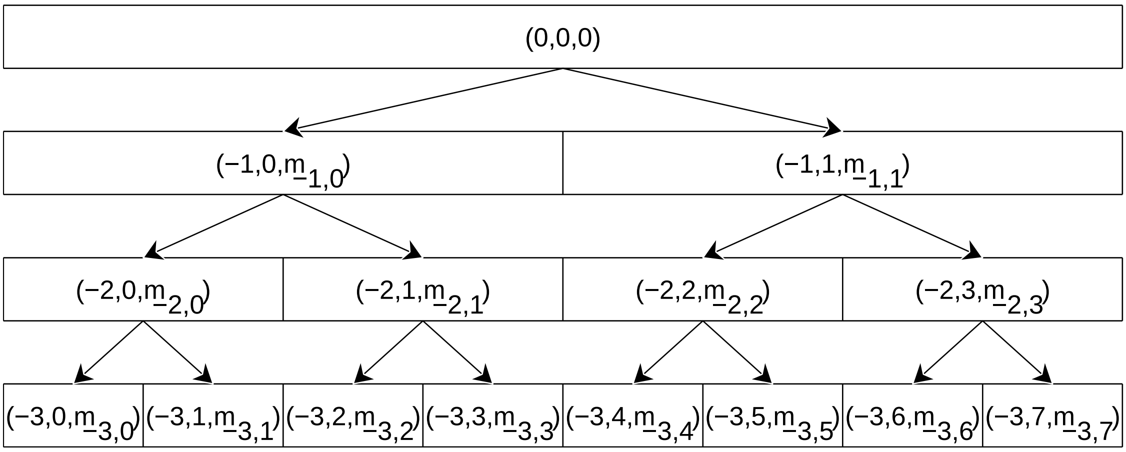


Figure 1: The extended set of wavelet packets organized in a binary tree structure. Each node in the tree is indexed by the triplet (ℓ, n, m) and represents the subspace $U_{\ell,n,m}$.

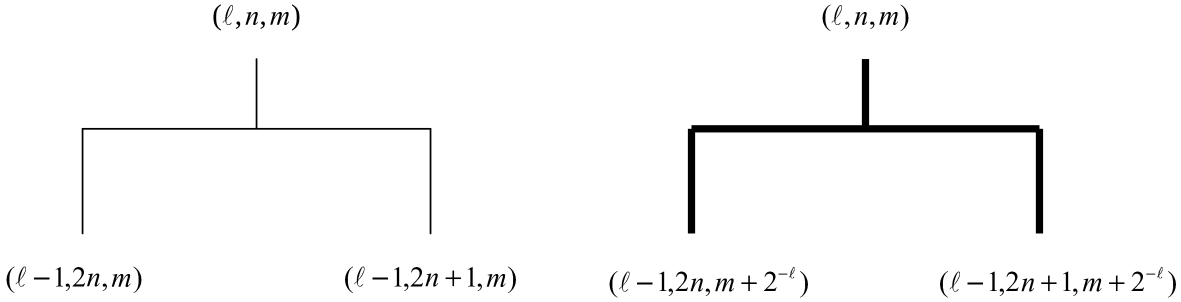


Figure 2: Alternative decompositions of a parent-node (ℓ, n, m) in a SIWPD tree. The branches to the children-nodes $(\ell - 1, 2n, m_c)$ and $(\ell - 1, 2n, m_c)$ are depicted by fine lines if $m_c = m$, and by heavy lines if $m_c = m + 2^{-\ell}$.

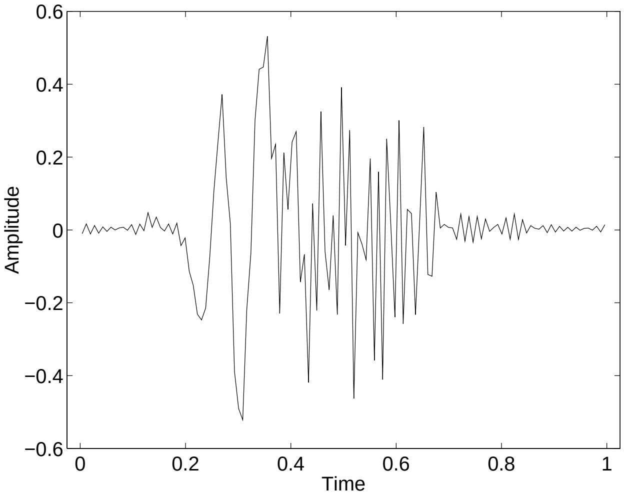
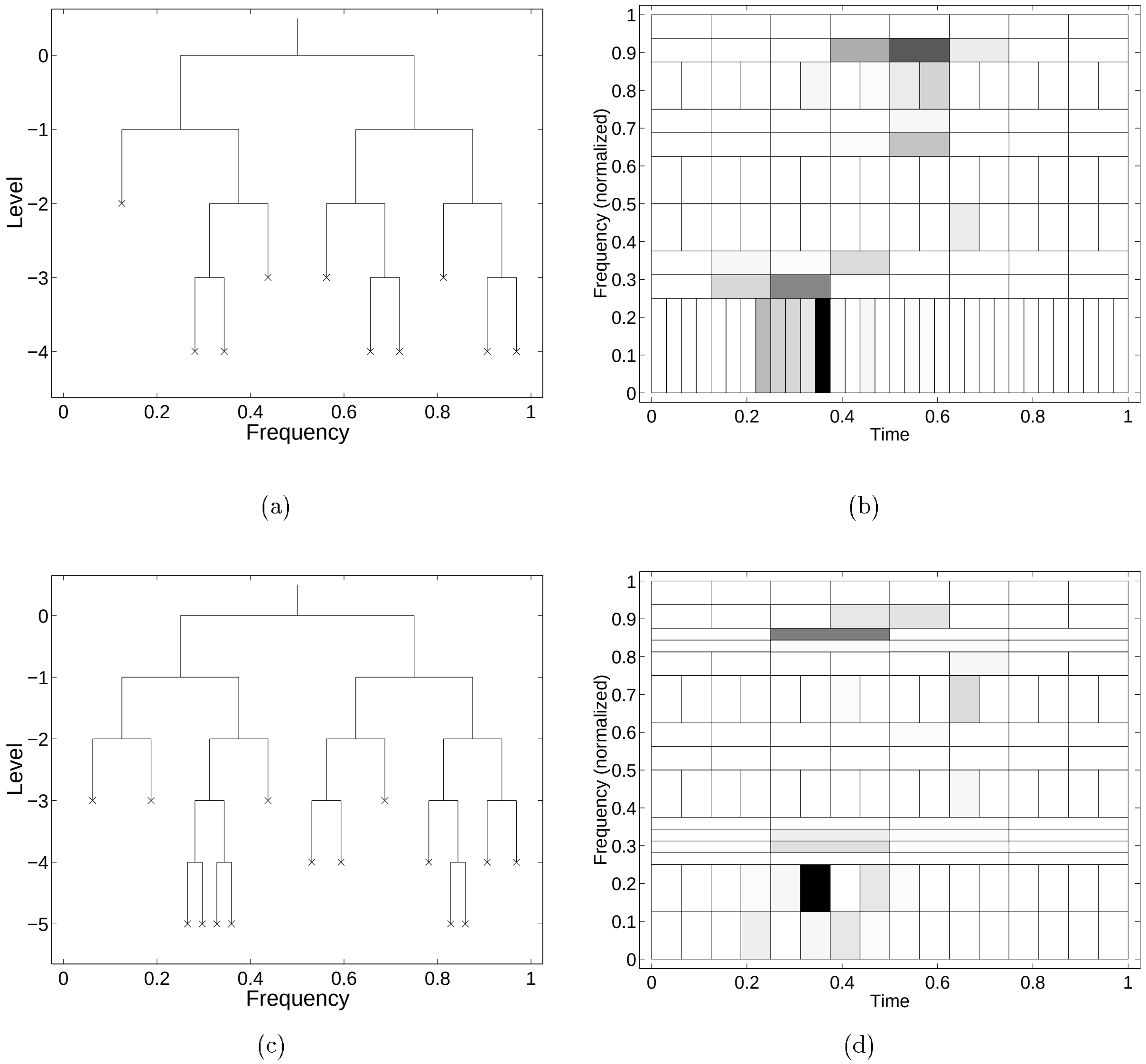


Figure 3: Test signal $g(t)$.



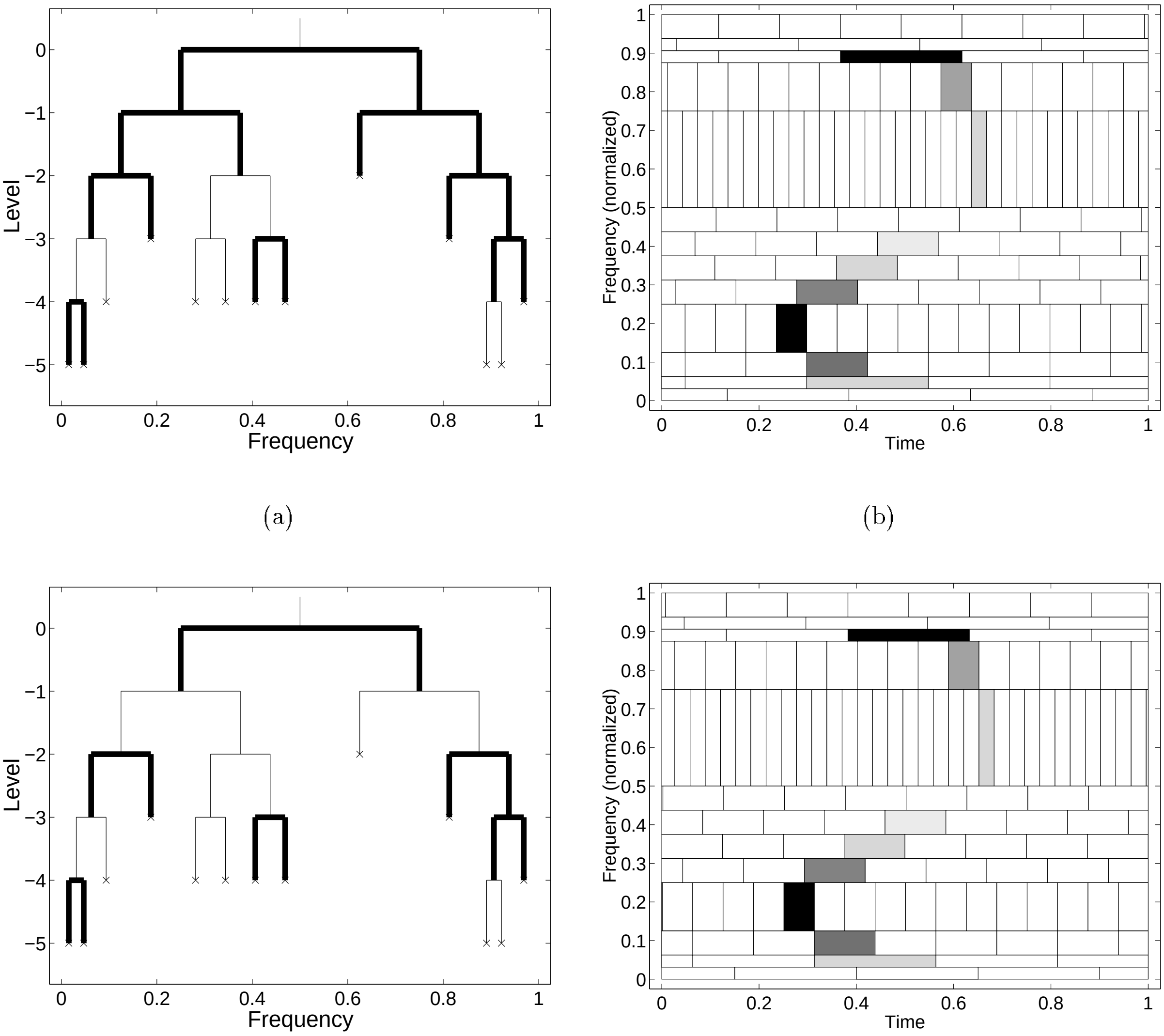
(a)

(b)

(c)

(d)

Figure 4: Effect of a temporal shift on the time-frequency representation using the WPD with 8-tap Daubechies least asymmetric wavelet filters: (a) The best expansion tree of $g(t)$. (b) $g(t)$ in its best basis; Entropy= 2.84. (c) The best expansion tree of $g(t-2^{-6})$. (d) $g(t-2^{-6})$ in its best basis; Entropy= 2.59.



(a)

(b)

(c)

(d)

Figure 5: Time-frequency representation using the SIWPD with 8-tap Daubechies least asymmetric wavelet filters: (a) The best expansion tree of $g(t)$. (b) $g(t)$ in its best basis; Entropy= 1.92. (c) The best expansion tree of $g(t-2^{-6})$. (d) $g(t-2^{-6})$ in its best basis; Entropy= 1.92. Fine and heavy lines in the expansion tree designate alternative node decompositions. Compared with the WPD (Fig. 4), beneficial properties are shift-invariance and lower information cost.

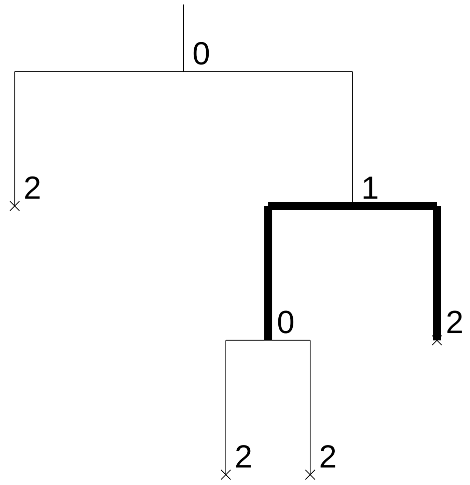
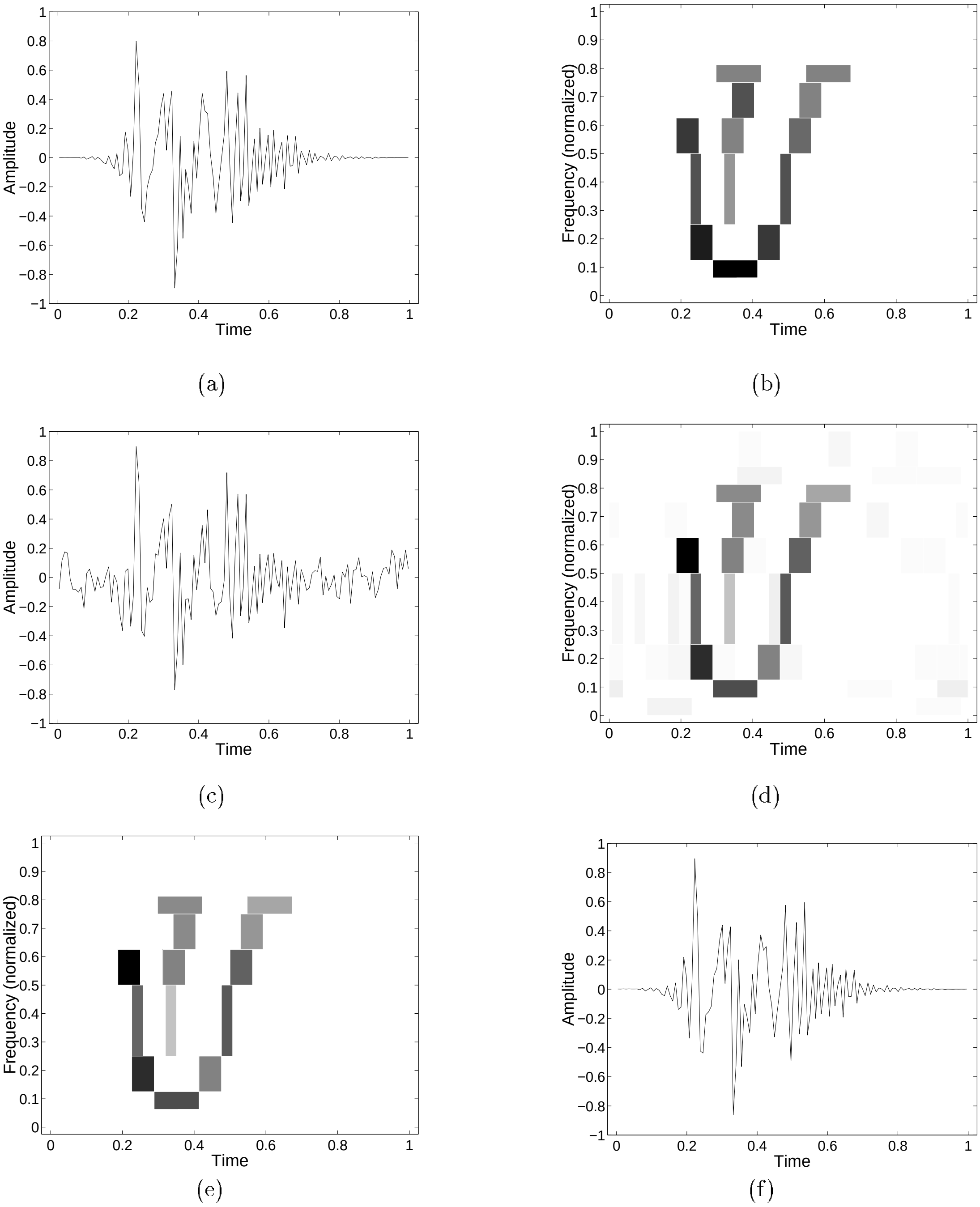
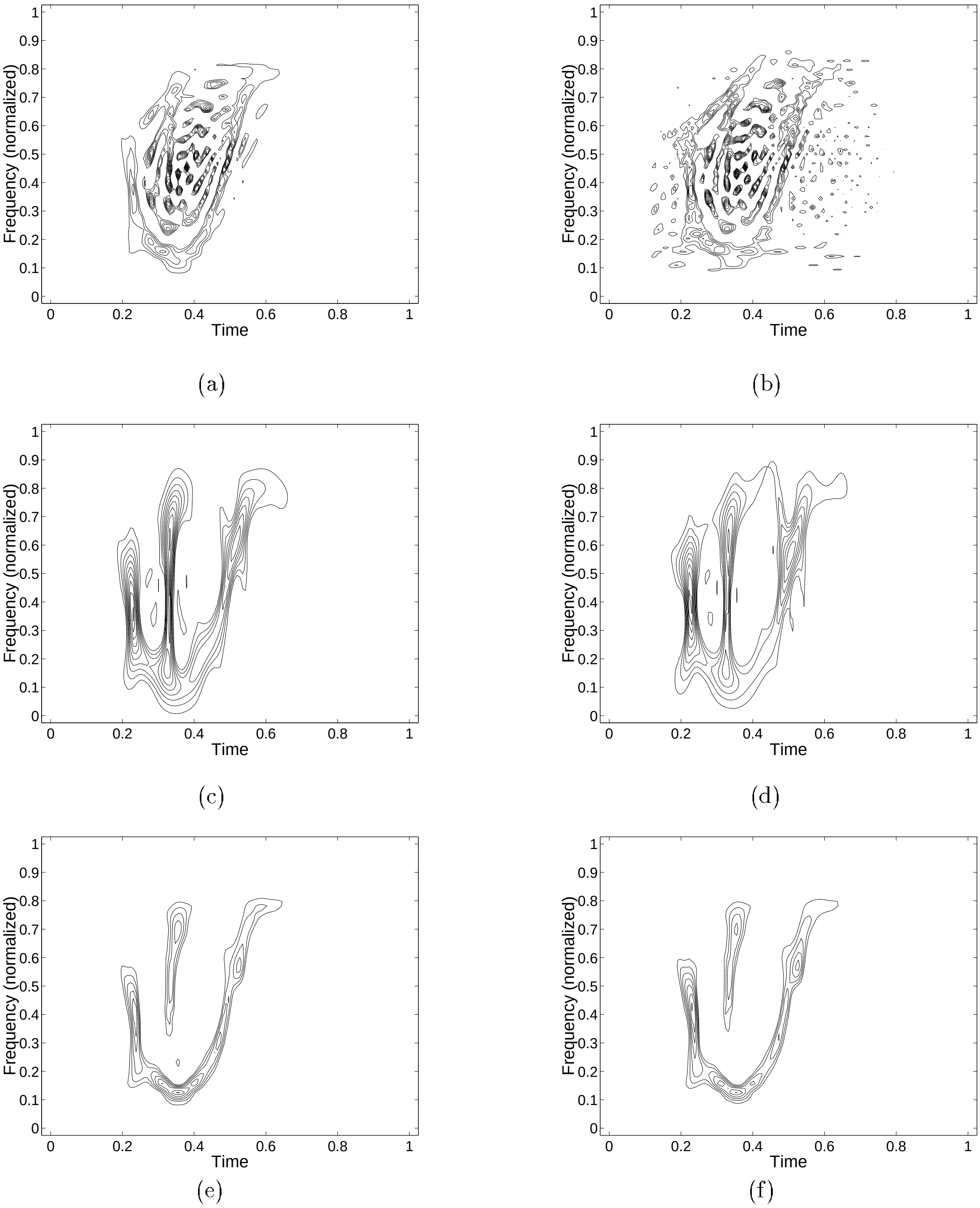


Figure 6: Exemplifying the description of SIWPD trees by 3-ary strings. Terminal nodes are represented by 2s, and internal nodes by either 0s or 1s, depending on their expansion mode. In the present example, the string is 0210222.



(a) (b) (c) (d) (e) (f)

Figure 7: Signal estimation by SIWPD and MDL principle: (a) Synthetic signal $f_1(t)$. (b) SIWPD of $f_1(t)$ using the Shannon entropy. (c) Noisy measurement $y_1(t)$; SNR= 7dB. (d) SIWPD of $y_1(t)$ using the MDL principle. (e) The expansion coefficients of $y_1(t)$ after hard-thresholding. (f) The signal estimate $\hat{f}_1(t)$; SNR= 19dB.



(a)

(b)

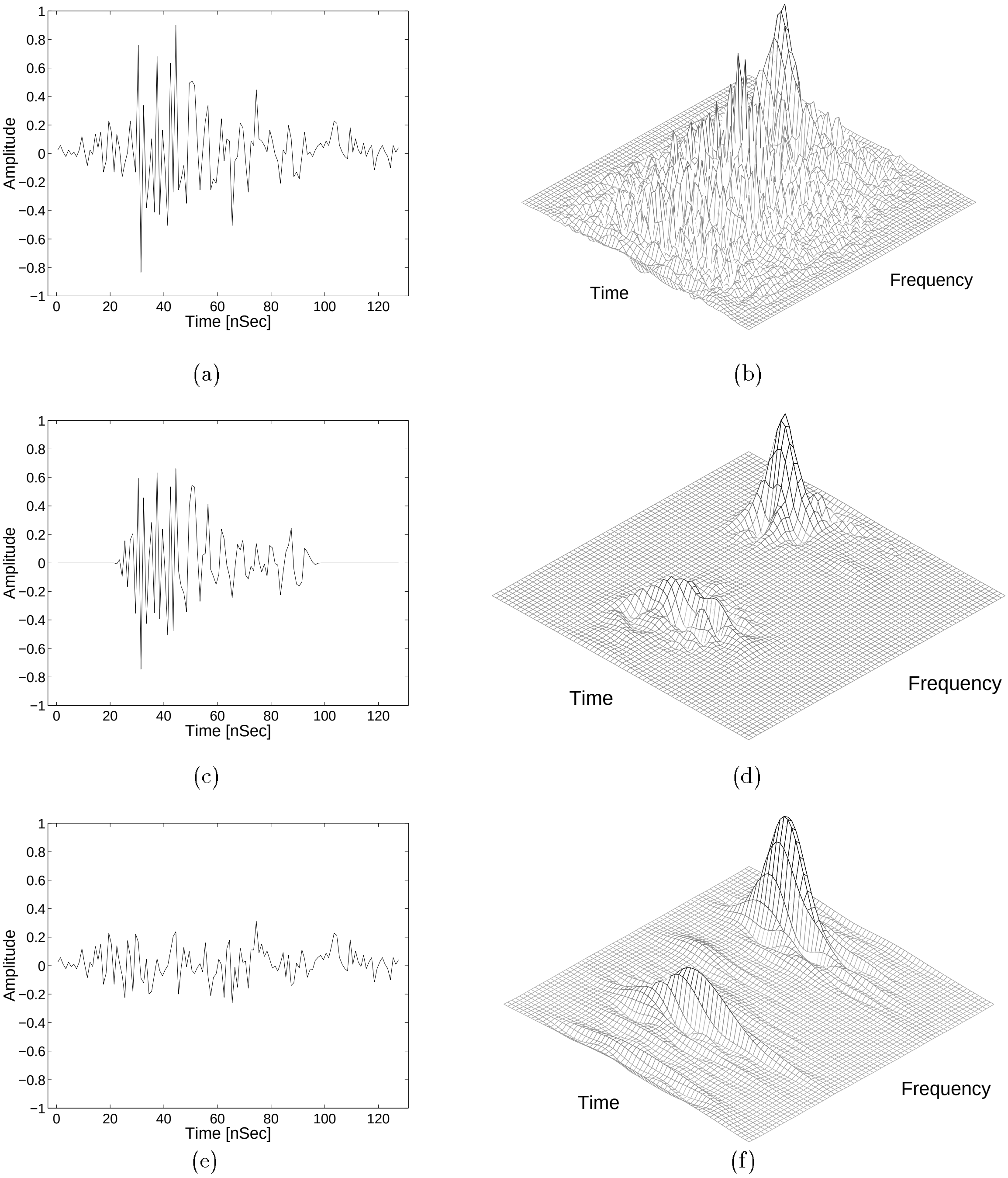
(c)

(d)

(e)

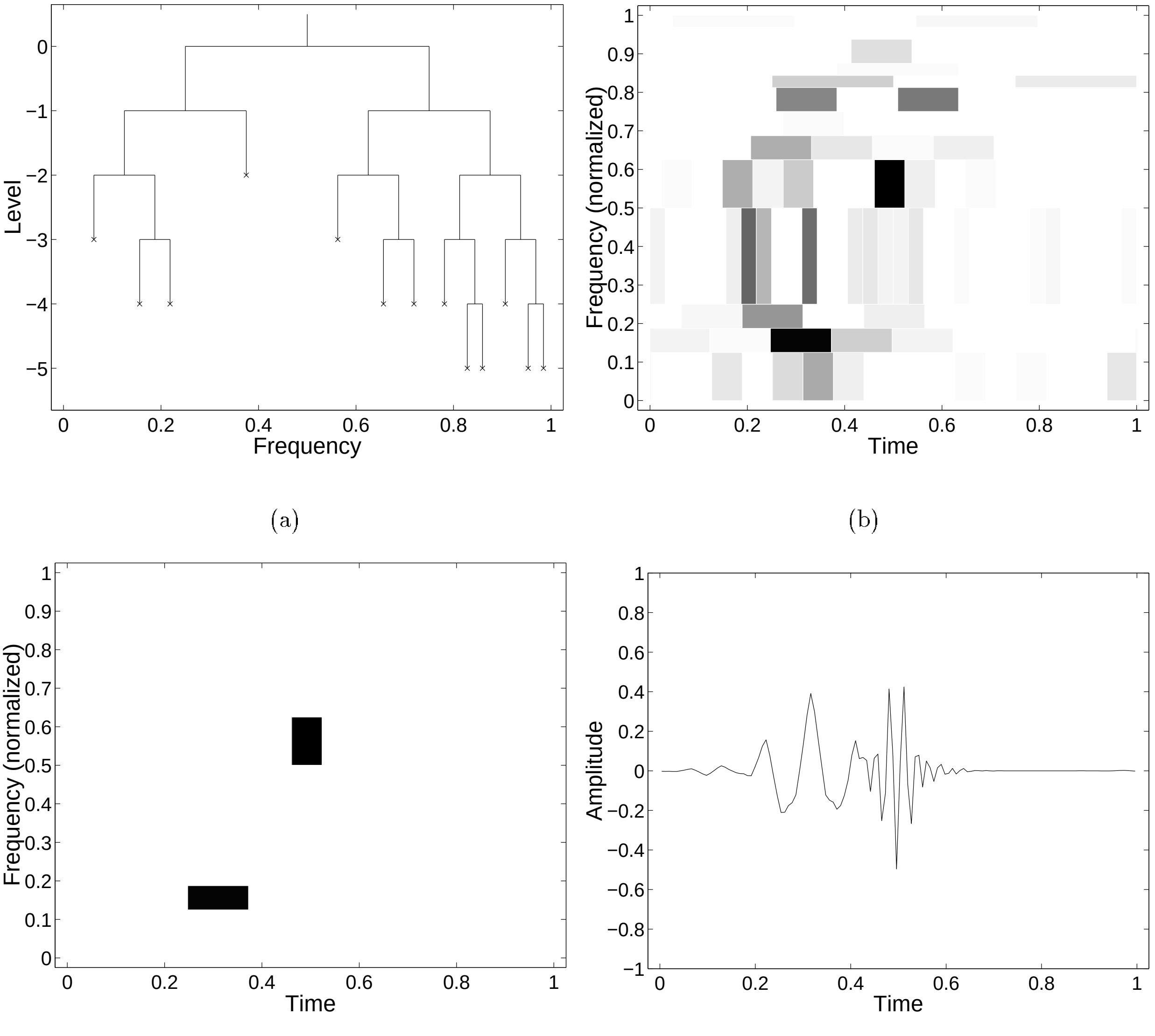
(f)

Figure 8: Contour plots of time-frequency distributions: (a) Wigner distribution for the original signal $f_1(t)$. (b) Wigner distribution for the noisy measurement $y_1(t)$. (c) Smoothed pseudo Wigner distribution for $f_1(t)$. (d) Smoothed pseudo Wigner distribution for $y_1(t)$. (e) The modified Wigner distribution for $f_1(t)$. (f) The estimate of the modified Wigner distribution for $y_1(t)$ by the MDL principle.



(a) (b) (c) (d) (e) (f)

Figure 9: Electromagnetic pulse in a relativistic magnetron (heterodyne detection; local oscillator= 2.6GHz): (a) Noisy measurement $y_2(t)$. (b) Wigner distribution for $y_2(t)$. (c) The signal estimate $\hat{f}_2(t)$ by the MDL principle. (d) The estimate of the modified Wigner distribution for $y_2(t)$. (e) Residual between $y_2(t)$ and $\hat{f}_2(t)$. (f) Smoothed pseudo Wigner distribution for $y_2(t)$.



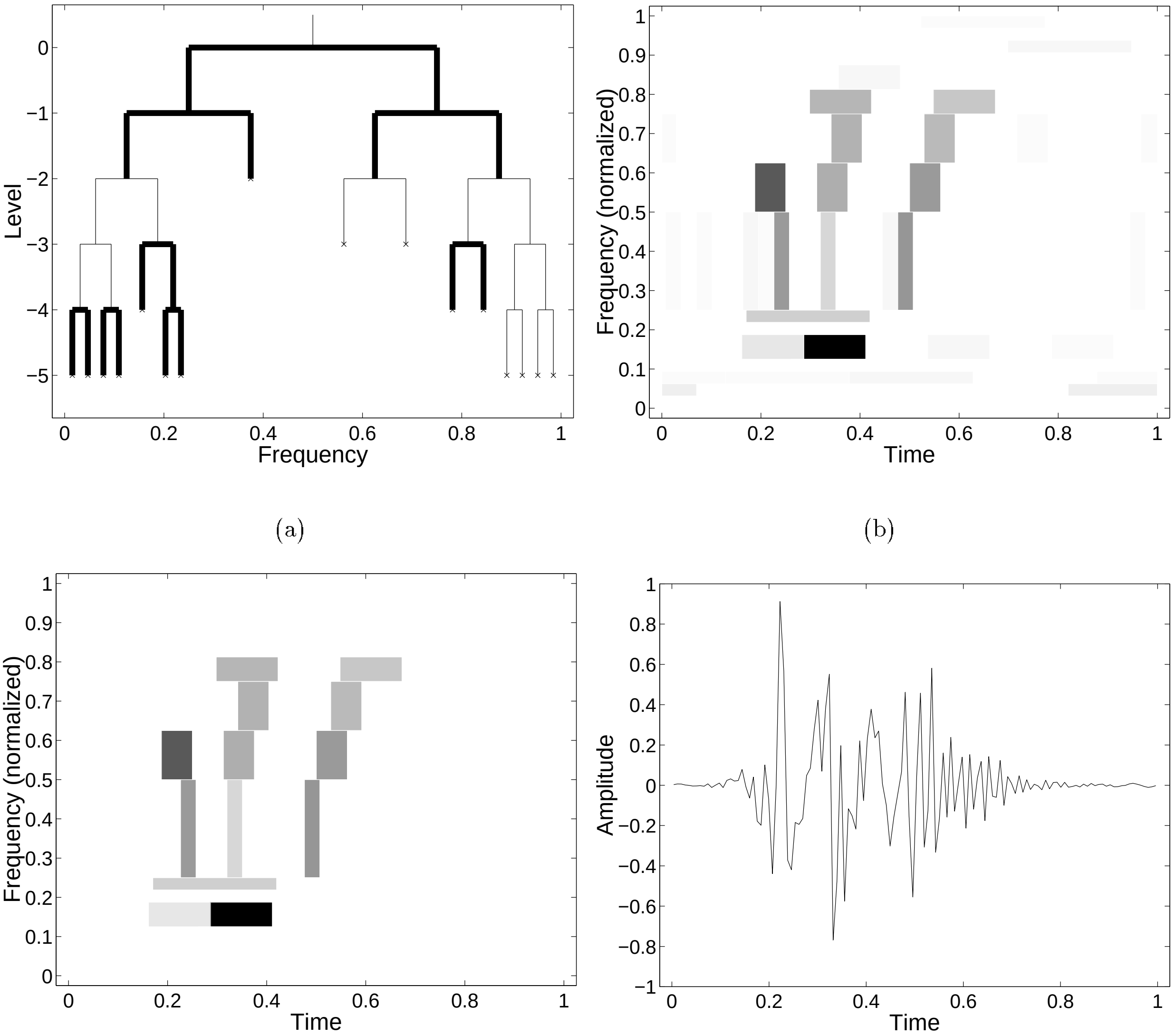
(a)

(b)

(c)

(d)

Figure 10: Signal estimation by the Saito method using the WPD: (a) The best expansion tree of $y_1(t)$ (the signal is depicted in Fig. 7(c)). (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 1.1dB.



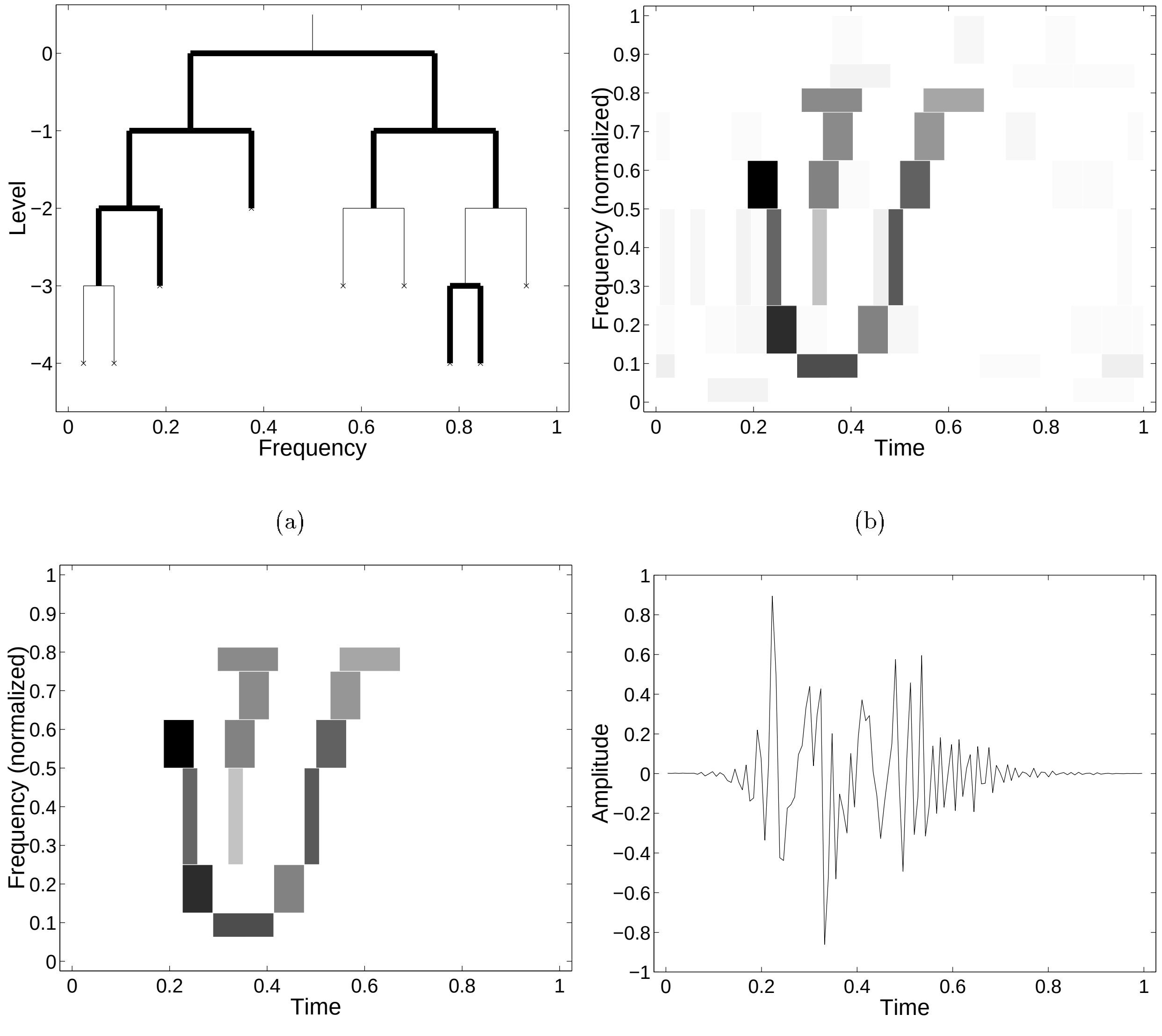
(a)

(b)

(c)

(d)

Figure 11: Signal estimation by the Saito method using the SIWPD: (a) The best expansion tree of $y_1(t)$. (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 12.8dB.



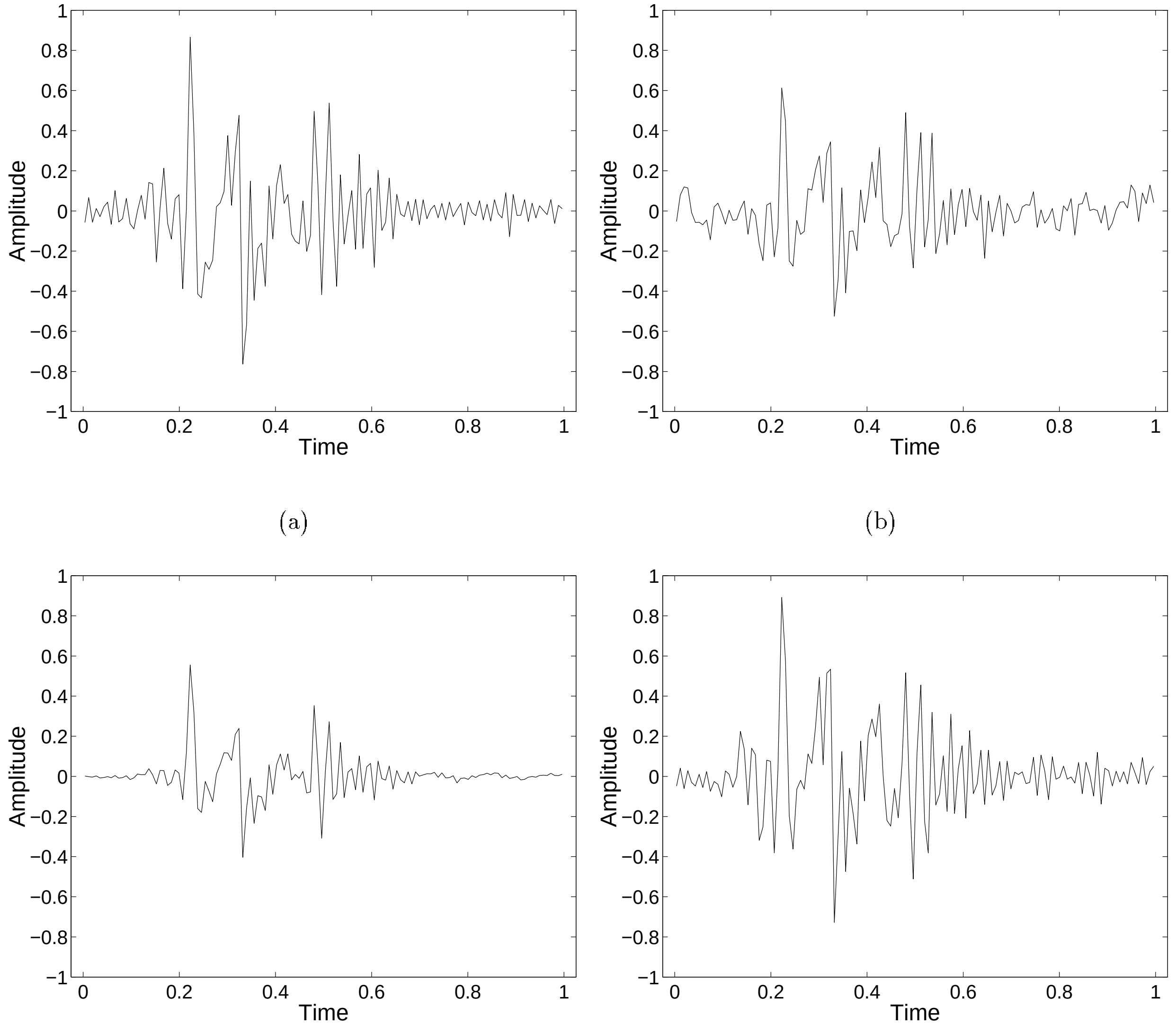
(a)

(b)

(c)

(d)

Figure 12: Signal estimation by the proposed method: (a) The optimal expansion tree of $y_1(t)$. (b) The expansion coefficients of $y_1(t)$. (c) The retained coefficients. (d) The signal estimate; SNR= 19dB.



(a)

(b)

(c)

(d)

Figure 13: Signal estimates of the synthetic signal using the library of wavelet packets (12-tap coiflet filters): (a) The Donoho-Johnstone method; SNR= 6.4dB. (b) The Method-of-Frames denoising (MOFDN); SNR= 7.1dB. (c) The Basis-Pursuit denoising (BPDN); SNR= 4.3dB. (d) The Matching-Pursuit denoising (MPDN); SNR= 7.5dB.