Signal processing via fast Malvar Wavelet transform algorithm

Eva Wesfreid*, M. Victor Wickerhauser+

*ENSTA, 32 Bd Victor, 75015 Paris, France, eva@ensta.fr, *CEREMADE, Université Paris-Dauphine-France.
+Department of Mathematics, Washington University, St. Louis, USA, victor@kirk.wustl.edu

RéSUMÉ

L'algorithme adaptatif de transformée en ondelettes de Malvar fournit une représentation spectrale complète et non redondante particulièrement adaptée pour l'analyse, la synthèse et la compression de signaux. Cet algorithme réalise une segmentation automatique du signal continu en unités quasi-stationnaires. En traitement de parole, on obtient une segmentation automatique en unités phonétiques; en outre, les centres de masse des fréquences associées à chaque unité phonétique ont été utilisés pour obtenir un algorithme de segmentation en parties voisées et non voisées.

ABSTRACT

The fast Malvar wavelet transform algorithm offers a complete and non redundant local spectrum representation which is useful for signal analysis, synthesis and compression. This algorithm performs an automatic segmentation of a continuous signal stream into quasi-stationary units. In speech processing this algorithm yields an automatic segmentation into phonetic units; furthermore, the frequencies center of mass associated to each phonetic unit is used to get a voiced/unvoiced segmentation algorithm.

1 Introduction

Malvar wavelet transform algorithm[12,30] consists in an arbitrary signal segmentation followed by a standard trigonometric transform (DCT, DST, ...) computed over preprocessed pieces[10,8] in order to eliminate redundancy and to preserve a complete signal description. A local spectrum representation over an arbitrary time partition is thus obtained.

An algorithm of entropy minimization yields a best time partition and an associated adapted local spectrum, it performs a signal segmentation in quasi-stationary units which appears to be useful in automatic recognition.

In speech processing, this algorithm splits a continuous stream into a sequence of quasi-stationary phonetic units.

In a previous paper[6], the local fundamental frequencies were used to realize a voiced unvoiced segmentation, further experiments showed that the frequencies center of mass of a voiced part is less than one eighth of the sampling rate, this threshold is used in this paper to distinguish a voiced part from an unvoiced one.

This paper is organized as follows: the Malvar wavelet transform algorithm is described in section 2 for any segmentation; in section 3 an entropy minimization algorithm allows us to select a signal segmentation and an associated adapted local spectrum; analysis, synthesis and compression is described in section 4; finally, the automatic segmentation of a speech continuous stream into phonetic units and into voiced/unvoiced parts is briefly described in section 5.

2 Malvar wavelet transform

Let us consider a real signal \( f(t) \in L^2(R) \), we shall compute the local spectrum associated to a Malvar wavelet transform over an arbitrary time-partition:

\[
R = \bigcup_{j \in \mathbb{Z}} I_j
\]

with \( I_j = [a_j, a_{j+1}] \), this local spectrum can be obtained via a standard fast trigonometric transform. We start with an arbitrary segmentation which is going to be preprocessed using a smooth rising cutoff function \( b_j(t) \) satisfying

\[
\begin{align*}
    b_j(t)^2 + b_j(2a_j - t)^2 &= 1 \\
    b_j(t) &= \begin{cases} 
        0 & \text{if } t < a_j - r \\
        1 & \text{if } t > a_j + r
    \end{cases}
\end{align*}
\]

with \( r > 0 \) such that \( a_j - r \leq a_{j+1} - r \) for \( 0 \leq j < N \). If

\[
\begin{cases} 
    \sin \frac{\pi}{T}(1 + \sin(\frac{\pi}{T}t)) & \text{if } -1 < t < 1 \\
    0 & \text{if } t < -1 \\
    1 & \text{if } t > 1
\end{cases}
\]

Figure 1: cutoff functions

then

\[
\begin{align*}
    b_j(t) &= \begin{cases} 
        b_j(t_a - t) & \text{if } a_j - r < t < a_j + r \\
        0 & \text{if } t < a_j - r \\
        1 & \text{if } t > a_j + r
    \end{cases}
\end{align*}
\]
with $b_j(t) \in C^1(t)$ and $b(2a_j - t) = b_j(t)$. The cutoff function is used to define the so-called folding operator \cite{folding}:

$$U_j f(t) = \begin{cases} b_j(t) f(t) + b(2a_j - t) f(2a_j - t) & \text{if } t \geq a_j \\ b_j(2a_j - t) f(t) - b_j(t) f(2a_j - t) & \text{if } t < a_j \end{cases} \quad (2)$$

and its adjoint unfolding operator:

$$U_j^* f(t) = \begin{cases} b_j(t) f(t) - b(2a_j - t) f(2a_j - t) & \text{if } t \geq a_j \\ b_j(2a_j - t) f(t) + b_j(t) f(2a_j - t) & \text{if } t < a_j \end{cases} \quad (3)$$

which satisfies

$$U_j^* U_j = U_j U_j^* = 1 \quad (4)$$

over $[a_j - r, a_j + r]$.

Observe that the folding operator splits $f(t)$ into

$$f_j(t) = \begin{cases} U_j f(t) & \text{if } t \in [a_j, a_j + r] \\ f(t) & \text{if } t \in [a_j + r, a_{j+1} - r] \\ U_{j+1}^* f(t) & \text{if } t \in [a_{j+1} - r, a_{j+1}] \end{cases} \quad (5)$$

Let us define

$$\phi_{j,k}(t) = \chi_{I_j}(t) g_{j,k}(t) \; (\chi_{I_j}(t) \text{ is equal to 1 if } t \in I_j \text{ and 0 otherwise}) \; \quad (6)$$

Since the system $\{\phi_{j,k}(t) : k \in \mathbb{N}\}$ forms an orthonormal basis for the set of functions $L^2(I_j)$ with polarities $(+,-)$ at $(a_j, a_{j+1})$, then

$$f_j(t) = \sum_{k \in \mathbb{N}} c_{j,k} \phi_{j,k}(t) \quad (7)$$

Had we chosen the other three pairs of signs in the folding operator definition, we would obtained three other trigonometric orthonormal basis: $\{\phi_{j,k}(t) : k \in \mathbb{N}, j \in \mathbb{Z}\}$

$$\phi_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \chi_{I_j}(t) \cos \frac{\pi}{|I_j|} (k + \frac{1}{2})(t - a_j) \quad (8)$$

where $k \in \mathbb{N}$, forms a local spectrum over $I_j$.

Furthermore, if

$$w_j(t) = \begin{cases} b_j(t) & \text{if } t \in [a_j - r, a_j + r] \\ 1 & \text{if } t \in [a_j + r, a_{j+1} - r] \\ b_j(2a_{j+1} - t) & \text{if } t \in [a_{j+1} - r, a_{j+1} + r] \end{cases} \quad (9)$$

denotes a window over $[a_j - r, a_{j+1} + r]$ and

$$\psi_{j,k}(t) = w_j(t) \phi_{j,k}(t)$$

then the local spectrum over $I_j$ can be represented via $\psi_{j,k}(t)$

$$c_{j,k} = <f_j(t), \phi_{j,k}(t)> = <f(t), \psi_{j,k}(t)> \quad (10)$$

for $k \in \mathbb{N}$ and $j \in \mathbb{Z}$. This result follows from the following property:

$$T_j \psi_{j,k}(t) = \begin{cases} \phi_{j,k}(t) & \text{if } t \in [a_j, a_j + r] \\ f(t) & \text{if } t \in [a_j + r, a_{j+1} - r] \\ U_{j+1}^* \psi_{j,k}(t) & \text{if } t \in [a_{j+1} - r, a_{j+1}] \end{cases} \quad (11)$$

for $j \in \mathbb{Z}$ and $k \in \mathbb{N}$.

The set of functions $\{\psi_{j,k}(t) : k \in \mathbb{N}, j \in \mathbb{Z}\}$ called Malvar Wavelets forms an orthonormal basis of $L^2(\mathbb{R})$, thus

$$\|f(t)\|^2 = \sum_{k \in \mathbb{N}} |c_{j,k}|^2 \quad (12)$$

Consequently, this signal decomposition into orthogonal trigonometric waveforms offers a complete and non redundant spectrum representation.

In the discrete case, the spectrum of the functions $f_j(t)$ with polarities $(+,-)$ at $(a_j, a_{j+1})$ can be computed via the standard fast DCT-IV transform algorithm \cite{fast_dct_4} over each $I_j$.

This fast DCT-IV transform algorithm can be then applied to the local spectrum over each $I_j$ to compute the functions $f_j$. The function $f(t)$ can be reconstructed from $\{f_j(t) : j \in \mathbb{Z}\}$ thanks to the unfolding operator defined in (3).

## 3 Entropy minimization algorithm

In this part, we describe an entropy minimization algorithm \cite{entropy_minimization} in order to select a adapted local spectrum.

Let us consider

- a sampled function $f$ over $[0, 2^N]$, 
- a time-partition for several levels $l = 0, 1, \ldots, \maxl$

$$[0, 2^N] = \bigcup_{0 \leq j < 2^l} I_j$$

where $I_j = [a_j^m, a_j^{m+1}]$ and $|I_j| = |a_j^{m+1} - a_j^m| = 2^{N-l}$

- a local spectrum

$$c_j^m = \{c_{j,k}^m : 0 \leq k < 2^{N-l}\}$$

computed over $I_j$ and

- the orthonormal basis

$$\{\psi_{j,k}^m : 0 \leq k < 2^{N-l}\}$$

Observe that $|I_j| = 2^{|I_j|}$ for $m = 1, 2, \ldots, \maxl$, $0 \leq j < 2^{\maxl-m}$, $0 \leq l < 2^{\maxl-m-1}$.
If $X_n^m$ denotes the space generated by $\{\psi_{n,k}^m : 0 \leq k < 2^{m-1}\}$ over $I_n^m$ then $f(t) \in X_n^m$ if and only if

$$f(t) = \sum_k c_{n,k}^m \psi_{n,k}^m(t)$$

and

$$X_n^m = X_{n+1}^{m-1} + X_{n+1}^{m-1}$$

Consequently, $X_n^m$ or $X_{n+1}^{m-1} + X_{n+1}^{m-1}$ can be chosen over $I_n^m = I_{n+1}^{m-1} \cup I_{n+1}^{m-1}$ using the following entropy function:

$$H(x) = \sum_x \frac{|a_x|^2}{|a_x|^2} \log \frac{|a_x|^2}{|a_x|^2}$$

for $x \in I^m$.

The entropy minimization algorithm is described in the following two steps:

**Step 0:**

Start with the local spectrum

$$s_0^m = c_0^m$$

($m = 0$, level $l = \text{maxl}$).

**Step 1:**

$$s_j^m = \begin{cases} c_j^m & \text{if } H(s_{j-1}^{m-1}) + H(s_{j-1}^{m+1}) > H(c_j^m) \\ c_j^m \cup c_{j+1}^m \cup c_{j-1}^m & \text{otherwise.} \end{cases}$$

(15)

for $m = 1, 2, \ldots, \text{maxl}$.

Let us consider $j \in [0, 512]$ in the following example (section 4) : since $H(c_j^0) + H(c_j^2) > H(c_j^1)$ and $H(c_j^2) + H(c_j^3) < H(c_j^1)$ then $s_0^j = c_j^0$ and $s_1^j = c_j^2 \cup c_j^0$. Thus the adapted local spectrum over $[0, 512]$ is $s_0^j = c_j^0 \cup c_j^2 \cup c_j^1$ because $H(s_0^j) + H(s_1^j) < H(c_j^0)$.

### 4 Analysis/synthesis, compression

Let us consider a speech signal sampled with a rate of about $8\text{KHz}$, corresponding to the first half second of the french sentence "des gens se sont levés dans les tribunes". Figure 2 shows the top 5% of the adapted local spectrum

$$s_{\text{maxl}}^m = c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0$$

(16)

(drawn in the middle) obtained via the entropy minimization algorithm when $N = 12$ and $\text{maxl} = 5$, its associated time partition

$$[0, 2^N] = I_0^0 \cup I_0^1 \cup I_0^2 \cup I_0^3 \cup I_0^4 \cup I_0^5 \cup I_0^6 \cup I_0^7 \cup I_0^8 \cup I_0^9 \cup I_0^{10} \cup I_0^{11} \cup I_0^{12} \cup I_0^{13}$$

(17)

is drawn with vertical lines. The smallest interval $I_0^7$ has been set to $16\text{ms}$ (128 samples).

![Figure 2: speech signal compression](image)

Figure 2 shows the original speech signal in the top part, the reconstructed signal obtained from the top 5% of the adapted local spectrum is drawn in its bottom part. Similar graphs are plotted by Xiang Fang to be used in [4] and [5]. Since the local spectrum $c_k^j$ over $I_0^7 = [0, 256]$ (samples) or $I_0^8 = [0, 32]$ (ms) is given by $c_k^j = \langle f_0^j(t), \phi_k^j(t) \rangle$ with

$$\phi_k^j(t) = x_h(t) \sqrt{2} \frac{\cos \pi (k + 1) t}{|I_0^7|}$$

and $k = 0, 1, \ldots, 256$, then the frequencies over $[0, 32]$ are $F_k = \frac{k + \frac{1}{2}}{|I_0^8|}$ and $0 \leq F_k < \frac{256 + \frac{1}{2}}{16 \times 32}$, consequently $0 \leq F_k < 4\text{KHz}$, because $\frac{32}{16}$ is the sample rate.

The local spectrum over $I_0^7 = [2048, 4096]$ (samples) ($I_0^8 = [256, 512]$ (ms)) is

$$c_k^j = \{ c_k^j \cup 0 < k < 2048 \}$$

with $c_k^j = \langle f_0^j(t), \phi_k^j(t) \rangle$

where

$$\phi_k^j(t) = x_h(t) \sqrt{2} \frac{\cos \pi (k + 1/2) t}{|I_0^7|}$$

with $k = 0, 1, \ldots, 2048$.

The frequencies over [2048, 4096] are $F_k = \frac{k + \frac{1}{2}}{|I_0^8|}$ and $0 \leq F_k < 4\text{KHz}$. The analysis, synthesis and compression in this example can be summarized as follows:

**Signal Analysis**

**Step 0**: Choose the smallest interval size $[|I_0^9| = 2^{N-\text{maxl}}]$, or equivalently the number of levels (maxl), $[|I_0^9| = 16\text{ms}]$ and $\text{maxl} = 5$.

**Step 1**: Signal preprocessing at each level ($l = 0, 1, \ldots, \text{maxl}$)

$$f(t) \mapsto \{ f_0^0(t), f_1^0(t), \ldots, f_{2^l}^0(t) \}$$

using the folding operator defined in (2).

**Step 2**: Compute a local spectrum at each level

$$\{ f_0^m(t), f_1^m(t), \ldots, f_{2^l}^m(t) \} \mapsto \{ c_0^m(t), c_1^m(t), \ldots, c_{2^l}^m(t) \}$$

using the fast DCT-IV transform.

**Step 3**: Select an adapted local spectrum

$$s_{\text{maxl}} = c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \cup c_3^0 \cup c_4^0 \ldots$$

using the entropy minimization algorithm (14), (15).

**Signal Synthesis**

**Step 4**: Reconstruct the preprocessed functions over the best time partition

$$s_{\text{maxl}} \mapsto \{ f_0^0(t), f_1^0(t), f_2^0(t), f_3^0(t), f_4^0(t), f_5^0(t), \ldots \}$$

using the fast DCT-IV transform.

**Step 5**: Reconstruct the original signal

$$\{ f_0^0(t), f_1^0(t), f_2^0(t), f_3^0(t), f_4^0(t), f_5^0(t), \ldots \} \mapsto f(t)$$

using the unfolding operator defined in (3).

**Compression**
The reconstructed signal was obtained with the top 5% of the spectral coefficients inside each interval of the best time partition, the other 95% has been cancelled:

\[ c_{jm}^n = c_{jm}^n = 0 \text{ if } |c_{jm}^n| < S_j \]

where

- \( S_j \) is \( \sigma_{jm}^n \)
- \( n \) is the integer part of \( |f_j^n| \times 5/100 \)
- \( \{ \sigma_{jm}^n \} \) is the decreasing sequence obtained from \( \{ c_{jm}^n \} \) via a sort function.

## 5 Speech processing

The frequencies center of mass

\[
CM[j] = \frac{\sum_k k c_{jk}^2}{\sum_k c_{jk}^2}
\]

of a voiced segment is less than one eighth of the sampling rate, this threshold (1 KHz our experiments) was used to get an automatic voiced/unvoiced segmentation.

![Figure 3: voiced/unvoiced segmentation](image)

The voiced/unvoiced segmentation of the speech signal used in the last section is represented in Figure 3.

The adapted Malvar wavelet algorithm decompose each signal into orthonormal trigonometric waveforms

\[
\psi_{jk}(t) = u_j(t) \frac{\sqrt{2}}{|I_j|} \cos \left( \frac{\pi}{I_j} (k + \frac{1}{2})(t - a_j) \right)
\]

whose duration \(|I_j|\) is variable. The shortest time lag can be chosen small enough \(|I_j| = 16\text{ms}\) in order to detect burst of plosive consonants\(^{60}\), rapid voicing onset of vowels and voiced-unvoiced segments. Other elementary waveforms representation can be found in [10], [11] and [12].

Due to the inertia of the vocal organs a new command may arrive before the preceding target is reached, our time-frequency representation offers a good description of this phenomenon of coarticulation. The entropy minimization algorithm yields a segmentation of a continuous speech stream into quasi-stationary phonetic units (Figure 2)

\[ f_{10}^0, f_{11}^1, f_{12}^2, f_{13}^3, f_{14}^4, f_{15}^5, \ldots \]

with its associated local spectrum

\[ c_0^1 \cup c_0^2 \cup c_0^3 \cup c_1^2 \cup c_1^3 \cup c_1^4 \cup c_1^5 \cup c_2^4 \cup c_2^5 \cup c_3^5 \ldots \]

- \( f_{10}^0 \) and \( f_{11}^1 \) represents the phonetic units of /d/ (\( f_{10}^0 \) is its burst).

### References


