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[1][2][3] consists in an arbitrary signal segmentation followed by a standard trigonometric transform (DCT, DST, ...) computed over preprocessed pieces[4][5] 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 :

$$R = \bigcup_{j \in Z} I_j$$

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

$$b_j(t)^2 + b_j(2a_j - t)^2 = 1$$

$$b_j(t) = \begin{cases} 0 & \text{if } t < a_j - r \\ 1 & \text{if } t > a_j + r \end{cases} \tag{1}$$

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

$$b(t) = \begin{cases} \sin \frac{\pi}{4}(1 + \sin(\frac{\pi}{2}t)) & \text{if } -1 < t < 1 \\ 0 & \text{if } t < -1 \\ 1 & \text{if } t > 1 \end{cases}$$

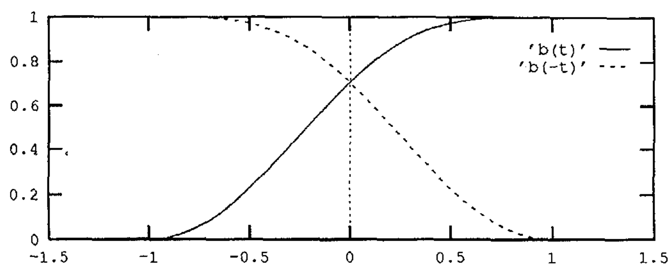

Figure 1: cutoff functions

then

$$b_j(t) = \begin{cases} b(\frac{t-a_j}{r}) & \text{if } a_j - r < t < a_j + r \\ 0 & \text{if } t < a_j - r \\ 1 & \text{if } t > a_j + r \end{cases}$$

with $b_j(t) \in C^1(t)$ and $b_j(2a_j - t) = b_j(t)$. The cutoff function is used to define the so-called *folding operator*[7] :

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

and its adjoint *unfolding operator* :

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

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 $\{ f_0(t), f_1(t), \ldots, f_j(t), \ldots, f_N(t)\}$ where

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

Let us define $\phi_{j,k}(t) = \chi_{I_j}(t) g_{j,k}(t)$ ($\chi_{I_j}(t)$ is equal to 1 if $t \in I_j$ and 0 otherwise) and

$$g_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \cos \frac{\pi}{|I_j|}(k + \frac{1}{2})(t - a_j) \quad (6)$$

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

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

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

$\phi_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \chi_{I_j}(t) \cos \frac{\pi}{|I_j|} k(t - a_j)$ if the polarities of $f_j(t)$ are $(+,+)$,

$\phi_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \chi_{I_j}(t) \sin \frac{\pi}{|I_j|}(k + \frac{1}{2})(t - a_j)\chi_{I_j}(t)$ if the polarities $f_j(t)$ are $(-,+)$,

$\phi_{j,k}(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \chi_{I_j}(t) \sin \frac{\pi}{|I_j|} k(t - a_j)$ if the polarities of $f_j(t)$ are $(-,-)$,

The following sequence of coefficients

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

where $k \in N$, forms a *local spectrum* over I_j.
Furthermore, if

$$w_j(t) = \begin{cases} b_j(t) & \text{if } t \in [a_j - r, a_j + r] \\ 1 & \text{if } t \in [a_j + r, a_{j+1} - r] \\ b_{j+1}(2a_{j+1} - t) & \text{if } t \in [a_{j+1} - r, a_{j+1} + r] \end{cases} \quad (7)$$

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

$$\psi_{j,k}(t) = w_j(t) g_{j,k}(t) \quad (8)$$

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

$$c_{j,k} = < f_j(t), \phi_{j,k}(t) > = < f(t), \psi_{j,k}(t) > \quad (9)$$

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

$$\begin{cases} \phi_{j,k}(t) = T_j \psi_{j,k}(t) \\ c_k = < f(t), \psi_{j,k}(t) > \end{cases} \quad (10)$$

where

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

for $j \in Z$ and $k \in N$.

The set of functions $\{\psi_{j,k}(t) : k \in N, j \in Z\}$ called *Malvar Wavelets* forms an orthonormal basis[1][2][3] of $L^2(R)$, thus

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

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*[8] 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 $\{f_j(t) : j \in Z\}$ thanks to the *unfolding* operator defined in (3).

3 Entropy minimization algorithm

In this part, we describe an *entropy minimization algorithm*[5] 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 $l = 0, 1, \ldots, maxl$

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

where $I_j^m = [a_j^m, a_{j+1}^m[$ and $|I_j^m| = |a_{j+1}^m - a_j^m| = 2^{N-l}$
- a local spectrum

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

computed over I_j^m and
- the orthonormal basis

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

Observe that $|I_j^m| = 2|I_i^{m-1}|$ for $m = 1, 2, \ldots, maxl$, $0 \leq j < 2^{lmaxl-m}$, $0 \leq i < 2^{lmaxl-m+1}$.

If X_j^m denotes the space generated by $\{\psi_{j,k}^m : 0 \leq k < 2^{N-l}\}$ over I_j^m then $f_j(t) \in X_j^m$ if and only if

$$f_j(t) = \sum_k c_{j,k}^m \psi_{j,k}^m(t)$$

and

$$X_j^m = X_{2j}^{m-1} + X_{2j+1}^{m-1}$$

Consequently, X_j^m or $X_{2j}^{m-1} + X_{2j+1}^{m-1}$ can be chosen over $I_j^m = I_j^{m-1} \cup I_{j+1}^{m-1}$ using the following entropy function :

$$H(x) = \sum_k \frac{|x_k|^2}{\|x\|^2} log \frac{|x_k|^2}{\|x\|^2} \quad (13)$$

for $x \in l^2$.

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

Step 0 :

We start with the local spectrum

$$s_j^0 = c_j^0 \quad (14)$$

($m = 0$, level $l = maxl$).

Step 1 :

$$s_j^m = \begin{cases} c_j^m & \text{if } H(s_{2j}^{m-1}) + H(s_{2j+1}^{m-1}) > H(c_j^m) \\ s_{2j}^{m-1} \cup s_{2j+1}^{m-1} & \text{otherwise.} \end{cases} \quad (15)$$

for $m = 1, 2, \ldots, maxl$.

Let us consider $j \in [0, 512]$ in the following example (section 4) : since $H(c_0^0) + H(c_1^0) > H(c_0^1)$ and $H(c_2^0) + H(c_3^0) < H(c_1^1)$ then $s_0^1 = c_0^1$ and $s_1^1 = c_2^0 \cup c_3^0$. Thus the *adapted local spectrum* over $[0, 512]$ is $s_0^2 = c_0^1 \cup c_2^0 \cup c_3^1$ because $H(s_0^1) + H(s_1^1) < H(c_0^2)$.

4 Analysis/synthesis, compression

Let us consider a speech signal sampled with a rate of about $8KHz$, 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*

$$s_0^{maxl} = c_0^1 \cup c_2^0 \cup c_3^0 \cup c_1^2 \cup c_2^2 \cup c_{13}^0 \cup c_{14}^0 \cup c_{15}^0 \ldots \quad (16)$$

(drawn in the middle) obained via the *entropy minimization algorithm* when $N = 12$ and $maxl = 5$, its associated *time partition*

$$[0, 2^N] = I_0^1 \cup I_2^0 \cup I_3^0 \cup I_1^2 \cup I_2^2 \cup I_{13}^0 \cup I_{14}^0 \cup I_{15}^0 \ldots \quad (17)$$

is drawn with vertical lines. The smallest interval I_j^0 has been set to $16ms$ ($128samples$).

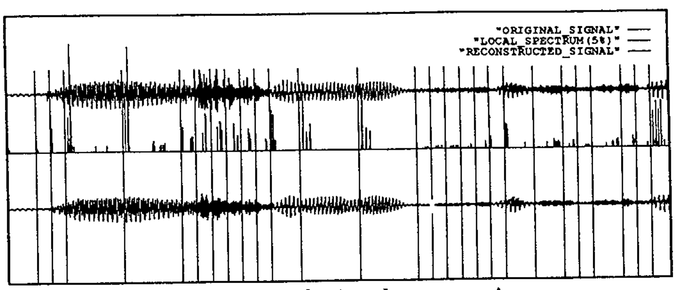

Figure 2: speech signal compression

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_0^1 over $I_0^1 = [0, 256]$ (samples) (or $I_0^1 = [0, 32]$ (ms)) is given by $c_{0,k}^1 =< f_0^1(t), \phi_0^1(t) >$

with

$$\phi_0^1(t) = \chi_{I_0^1}(t) \frac{\sqrt{2}}{\sqrt{|I_0^1|}} \cos \frac{\pi}{|I_0^1|}(k + \frac{1}{2})t$$

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

The local spectrum over $I_1^4 = [2048, 4096]$ (samples) ($I_1^4 = [256, 512]$ (ms)) is

$$c_1^4 = \{c_{1,k}^4 : 0 \leq k < 2048\}$$

with $c_1^4 =< f_1^4(t), \phi_1^4(t) >$

where

$$\phi_1^4(t) = \chi_{I_1^4}(t) \frac{\sqrt{2}}{\sqrt{|I_1^4|}} \cos \frac{\pi}{|I_1^4|}(k + \frac{1}{2})(t - 2048)$$

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

The frequencies over $[2048, 4096]$ are $F_k = \frac{k+\frac{1}{2}}{|2I_1^4|}$ and $0 < F_k < 4KHz$. The analysis, synthesis and compression in this example can be summarized as follows :

Signal Analysis

Step 0 : Choose the smallest interval size ($|I_j^0| = 2^{N-maxl}$) or equivalently the number of levels ($maxl$), ($|I_j^0| = 16ms$ and $maxl = 5$).

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

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

using the *folding* operator defined in (2).

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

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

using the fast DCT-IV transform.

Step 3 : Select an *adapted local spectrum*

$$s_0^{maxl} = c_0^1 \cup c_2^0 \cup c_3^0 \cup c_1^2 \cup c_2^2 \cup c_{13}^0 \cup c_{14}^0 \cup c_{15}^0 \ldots$$

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

Signal Synthesis

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

$$s_0^{maxl} \longmapsto \{f_0^1, f_2^0, f_3^0, f_1^2, f_2^2, f_{13}^0, f_{14}^0, f_{15}^0, \ldots\}$$

using the fast DCT-IV transform.

Step 5 : Reconstruct the original signal

$$\{f_0^1, f_2^0, f_3^0, f_1^2, f_2^2, f_{13}^0, f_{14}^0, f_{15}^0, \ldots\} \longmapsto f(t)$$

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 :

$$\{c^m_{j,k} : c^m_{j,k} = 0 \text{ if } |c^m_{j,k}| < S_j\}$$

where

- $S_j = d^m_{j,a}$,
- a is the integer part of $|I^m_j| * 5/100$,
- $\{d^m_j\}$ is the decreasing sequence obtained from $\{c^m_j\}$ via a sort function.

5 Speech processing

The frequencies center of mass

$$CM[j] = \frac{\sum_k k c^2_{j,k}}{\sum_k c^2_k} \tag{18}$$

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.

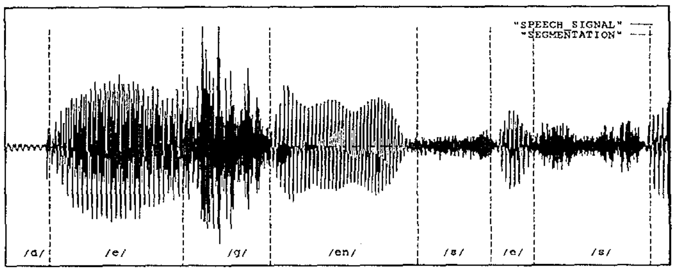

Figure 3: voiced/unvoiced segmentation

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*

$$\psi_{j,k}(t) = w_j(t)\frac{\sqrt{2}}{\sqrt{|I_j|}} \cos \frac{\pi}{|I_j|}(k+\frac{1}{2})(t-a_j) \tag{19}$$

whose duration $|I_j|$ is variable. The shortest time lag can be chosen small enough ($|I^0_j| = 16ms$) in order to detecte burst of plosive consonants[9], 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)

$$f^1_0, f^0_2, f^0_3, f^2_1, f^2_2, f^0_{13}, f^0_{14}, f^0_{15}, \ldots$$

with its associated local spectrum

$$c^1_0 \cup c^0_2 \cup c^0_3 \cup c^2_1 \cup c^2_2 \cup c^0_{13} \cup c^0_{14} \cup c^0_{15}, \ldots$$

- f^1_0 and f^0_2 represents the phonetic units of /d/ (f^0_2) is its burst),
- $f^0_3, f^2_1, f^2_2, f^0_{13}$ represents the phonetic units of /e/ (f^0_3 can be seen as the attack and f^0_{13} as the decay), f^0_{13} is a mixed voiced/unvoiced unit, it corresponds also to the begining of consonant /g/.

References

[1] R. R. Coifman and Y. Meyer, *Remarques sur l'analyse de Fourier à fenêtre*, série I, C. R. Acad. Sci. Paris **312**, pp. 259-261. (1991)

[2] H. Malvar, *Lapped transforms for efficient transform/subband coding*, IEEE Trans. Acoustics, Speech and Signal Processing, **38**, pp. 969-978.(1990)

[3] H. Malvar, *Signal Processing with Lapped transforms*, ARTECH HOUSE, Boston, London (1992)

[4] V. Wickerhauser, INRIA Lecture on Wavelet Packet Algorithms, pp. 21-28.(1991)

[5] R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best-basis selection*, IEEE Trans.Info.Theory (March, 1992).

[6] E. Wesfreid and M. V. Wickerhauser, *Adapted trigonometric transform and speech processing*, IEEE Trans. Acoustic. Speech Processing (special wavelets), Dec 1993 (to appear).

[7] P. Auscher, G. Weiss and M. V. Wickerhauser, *Local cosine Transform* in *Wavelets and Their Applications*, Wavelets : A Tutorial in Theory & Applications Ed, by C. K. Chui (Acdemic Press) 1992.

[8] Yip Rao, *Discrete cosine transform* (1989)

[9] Calliope , *La parole et son traitement automatique* (Masson - Paris) (1989)

[10] C. D'Alessandro and X.Rodet, *Synthèse et analyse - synthèse par fonction d'ondes formantiques*, J. Acoustique **2**, pp. 163-169.(1989)

[11] J. S. Liénard, *Speech analysis and reconstruction using short time, elementary waveforms*, Proc. IEEE ICASSP-87, Dallas, pp. 948-951.(1987)

[12] X.Rodet, *Time domain formant-wave-function synthesis*, in : *Spoken Language Generation and Understanding*, ed. by J. C. Simon (D. Reidel Publishing Compagny - Dordrecht Holland)(1980).