Adapted Local Trigonometric Transforms and Speech Processing

Eva Wesfreid[1]
Mladen Victor Wickerhauser[2*]

[1]CEREMADE, Université Paris IX-Dauphine
[2]Department of Mathematics, Washington University in St. Louis

1 Introduction

In this note, we apply an algorithm, based on the local trigonometric orthonormal basis and the adapted local trigonometric transform, to decompose digitized speech signals into orthogonal elementary waveforms. This algorithm leads to a local time-frequency representation which is well adapted to analysis-synthesis, compression and segmentation. We present some applications and experimental results for signal compression and automatic voiced-unvoiced segmentation. Furthermore, compression provides a simplified decomposition which appears to be useful for detecting fundamental frequencies and characterizing formants.

We begin with a clean, digitized speech signal. The signal is decomposed into a local trigonometric orthonormal basis which consists of cosines or sines multiplied by smooth cutoff functions. This basis is described by R. Coifman and Y. Meyer [3] and by H. Malvar [7]. We describe and then apply to speech processing an adapted version of this lapped orthogonal transform, which includes a "best basis method" obtained by entropy minimization [4]. This algorithm produces an adapted orthogonal elementary waveform decomposition, which is a local spectral representation for the speech signal. Roughly speaking, we get a windowed cosine transform of the signal, with the window size well adapted to the spectrum it contains. The associated time-partition, or choice of windows, appears to be useful for segmentation into voiced and unvoiced portions, which can be recognized by the number of peaks or "theoretical dimension" of the local spectrum. The time-partition provides short segments where there is fast frequency variation and long segments where there is slow frequency variation. The spectral representation is invertible and allows both perfect reconstruction (analysis-synthesis) and lossy approximation (compression).

We introduce a *formant representation* as follows. We examine the spectrum in each segment and locate the centers of mass for the top few peaks. Keeping just the top few peaks, or just a few of the most energetic waveform components, is a kind of compression, and computing the centers of mass of the peaks is a drastic reduction of the amount of data to be used for subsequent recognition. The formant representation is the resulting set of locally constant spectral lines, or step function approximations to the time-frequency function.

[*] Research partially supported by AFOSR Grant F49620-92-J-0106

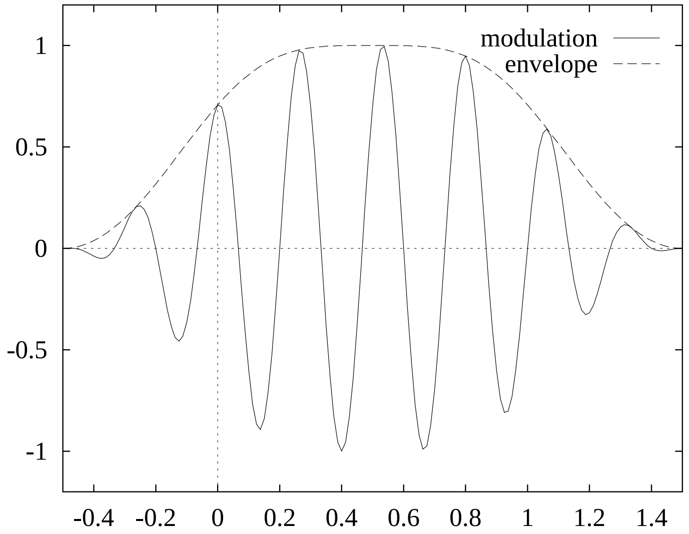

Figure 1: A basis function for the lapped orthogonal transform

For comparison and to suggest variations of this algorithm, we point out that other elementary waveform representations can be used in similar adapted decompositions and segmentations. Some of these are described in references [5], [6] and [9].

2 Algorithm description

We recall briefly the local cosine transform [1]. Let us consider a partition of the line:

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

with $I_j = [a_j, a_{j+1})$, such that the width of the intervals is never less than a fixed positive number: $a_{j+1} - a_j \geq \epsilon > 0$ for all $j \in Z$. We define the following cutoff functions:

$$b_j(t) = \begin{cases} \beta(\frac{t-a_j}{r}) & t \in [a_j - r, a_j + r) \\ 1 & t \in [a_j + r, a_{j+1} - r) \\ \beta(\frac{a_{j+1}-t}{r}) & t \in [a_{j+1} - r, a_{j+1} + r) \\ 0 & t \in (-\infty, a_j - r] \cup [a_{j+1} + r, \infty) \end{cases}$$

with $\beta(t) = \sin[\frac{\pi}{4}(1 + \sin\frac{\pi}{2}t)]$ and $0 < r \leq \epsilon$. This function, adjusted to the interval $[0, 1]$, is the envelope of the curve displayed in Fig.1. The set of functions:

$$\Psi_k^j(t) = b_j(t)\frac{\sqrt{2}}{\sqrt{|I_j|}}\cos\frac{\pi}{|I_j|}(k + \frac{1}{2})(t - a_j),$$

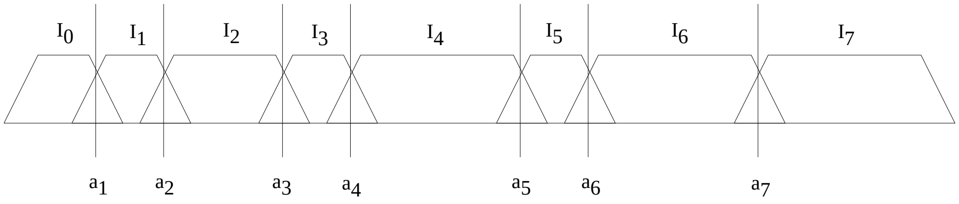

Figure 2: Lapped orthogonal basis functions on adjacent intervals

with $j \in Z$ and $k \in N$, is an orthonormal basis for $L^2(R)$. Consequently, each signal $S(t) \in L^2(R)$ can be written in terms of the functions Ψ_k^j:

$$S(t) = \sum_{\substack{j \in Z \\ k \in N}} c_k^j \Psi_k^j(t)$$

with

$$c_k^j = \langle S(t), \Psi_k^j(t) \rangle = \frac{\sqrt{2}}{\sqrt{|I_j|}} \int S(t) b_j(t) \cos \frac{\pi}{|I_j|}(k + \frac{1}{2})(t - a_j) \; dt \tag{1}$$

A superposition of these functions may be depicted by a sequence of adjacent envelopes or windows, with vertical lines drawn between the nominal window boundaries. This is done in Fig.2.

It is possible to compute several local cosine transforms all at once, recursively subdividing the intervals into halves. The basis functions on each subinterval are the orthogonal direct sum of the basis functions on its left and right halves, and this orthogonality propagates up through the multiple levels of the binary "family tree" in Fig.3.

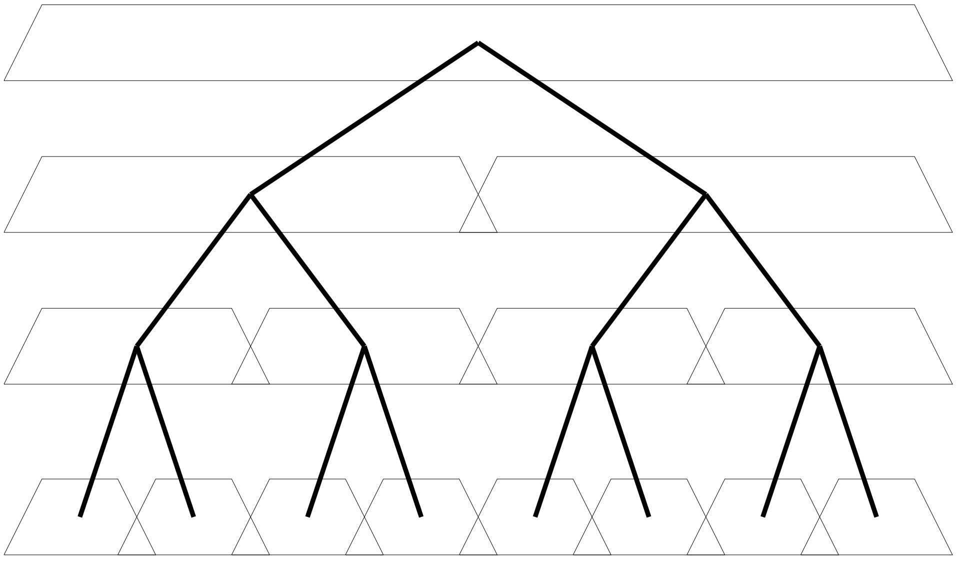

Figure 3: Several lapped orthogonal transforms computed all at once

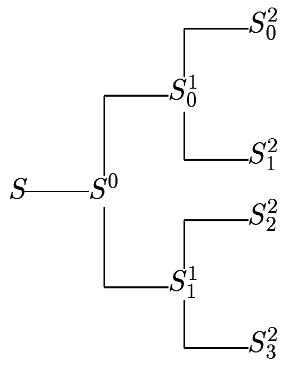

Figure 4: 3-level decomposition into segments by “folding” and restriction

The inner products in Eq.(1) can be computed using a standard fast discrete cosine transform, after a reliminary “folding” step described in [10]. This “folding” splits S(t) into a set of local finite energy signals $S_j(t) \in L^2(I_j)$, $j \in Z$, such that applying a standard discrete cosine transform to the coefficients in $S_j(t)$ is equivalent to computing all inner products with the functions Ψ_k^j. In other words,

$$S_j(t) = \sum_{k \in Z} c_k^j \; \phi_k^j(t) \tag{2}$$

with

$$c_k^j = \langle S_j(t), \phi_k^j(t) \rangle \tag{3}$$

where

$$\phi_k^j(t) = \frac{\sqrt{2}}{\sqrt{|I_j|}} \cos \frac{\pi}{|I_j|}(k + \frac{1}{2})(t - a_j)\chi_{I_j}(t) \tag{4}$$

Discrete sampled cosines at half-integer frequencies are the basis functions of the “DCT-IV” transform. The characteristic function $\chi_{I_j}(t)$ is equal to 1 if $t \in I_j$ and 0 otherwise. If $S(t)$ is a sampled signal with $t \in \{0, 1, 2, \ldots, 2^N - 1\}$, then we can “fold” first at the boundaries which gives $S^0(t)$ and next recursively at the middle points in a few levels. Therefore, this “folding” splits each function $S_j^\ell(t) \in L^2(I_j^\ell)$ into $S_{2j}^{\ell+1}(t) \in L^2(I_{2j}^{\ell+1})$ and $S_{2j+1}^{\ell+1}(t) \in L^2(I_{2j+1}^{\ell+1})$ as seen in Fig.4. We can calculate the standard DCT-IV transform [8] for each S_j^ℓ which gives the spectral tree in Fig.5.

The orthogonality of the lapped orthogonal transform implies the following energy conservation identities:

$$\|S\|^2 = \|S^0\|^2 = \|S^j\|^2 \overset{\text{def}}{=} \sum_k \|S_k^j\|^2; \qquad \|S_k^j\|^2 = \|d_k^j\|^2$$

If $\{x_k\}$ belongs to l^2 and $l^2 \log l^2$ then we can define the *spectral entropy* of $\{x_k\}$ to be

$$H(x) = -\sum_k \frac{|x_k|^2}{\|x\|^2} \log \frac{|x_k|^2}{\|x\|^2} = \frac{\lambda(x)}{\|x\|^2} + \log \|x\|^2, \tag{5}$$

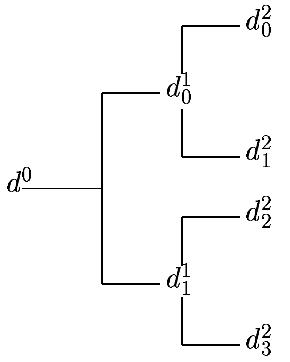

Figure 5: 3-level decomposition into local cosine transforms

with $\lambda(x) = -\sum_k |x_k|^2 \log |x_k|^2$. Then $\exp(H(x)) = \|x\|^2 \exp(\frac{\lambda(x)}{\|x\|^2})$ may be called the *theoretical dimension* of the sequence $\{x\}$.

The *adapted local spectrum* a^ℓ_j over the time interval I^ℓ_j is defined recursively in terms of the spectra on subintervals of I^ℓ_j, by the following formula:

$$a^\ell_j = \begin{cases} d^\ell_j & \text{if} \quad H(d^\ell_j) < H(a^{\ell+1}_{2j}) + H(a^{\ell+1}_{2j+1}), \\ a^{\ell+1}_{2j} \cup a^{\ell+1}_{2j+1} & \text{otherwise.} \end{cases}$$

See [4] for a detailed description of this algorithm. If we begin with $a^\ell_M = d^\ell_M$ at some deepest decomposition level M, then a^0_0 will contain the lapped orthogonal transform coefficients with respect to that division into intervals I^ℓ_j which minimizes the total spectral entropy. We will call the chosen division into intervals the *adapted (entropy-minimizing) time-partition* for the given signal.

3 Applications to Speech Processing

In this section we suppose that our signal $S = S(t)$ is speech sampled over a total time interval $[0, T]$, and that we have calculated its adapted time-partition:

$$[0, T] = \bigcup_{0 \le j < N} I_j.$$

The "folding" of S at the boundaries between subintervals can be viewed as a segmentation of the function:

$$S(t) \longrightarrow (S_0(t), S_1(t), \ldots, S_{N-1}(t)).$$

Using Eqs.(2), (3) and (4) each $S_j(t)$ can be decomposed into a set of orthogonal elementary waveforms:

$$S_j(t) = \sum_{0 \le k < n_j} c^j_k \; \phi^j_k(t),$$

with n_j as number of samples in I_j and with c_k^j as spectral coefficients computed via the standard DCT-IV transform [8]. The analysis-synthesis may be represented by the following scheme:

$$\begin{tikzcd}
 & S_0(t) \arrow[r, "\text{DCT-IV}"] & c_k^0 \arrow[r, "\text{DCT-IV}"] & S_0(t) & \\
 & S_1(t) \arrow[r, "\text{DCT-IV}"] & c_k^1 \arrow[r, "\text{DCT-IV}"] & S_1(t) & \\
S(t) \arrow[r, "\textit{fold}"] & \vdots & \vdots & \vdots \arrow[r, "\textit{unfold}"] & S(t). \\
 & S_{N-1}(t) \arrow[r, "\text{DCT-IV}"] & c_k^{N-1} \arrow[r, "\text{DCT-IV}"] & S_{N-1}(t) &
\end{tikzcd}$$

Each discrete local cosine coefficient c_k^j gives the amplitude of an elementary orthogonal waveform component. This elementary waveform's period is $T_k = \frac{2\pi}{\omega_k}$ with $\omega_k = \frac{\pi}{|I_j|}(k + \frac{1}{2})$, therefore its frequency is

$$F_k = \frac{\omega_k}{2\pi} = \frac{k + \frac{1}{2}}{2|I_j|} \tag{6}$$

The maximum frequency that we can distinguish is about half the sampling rate $n_j/|I_j|$, and thus

$$0 \leq F_k < \frac{n_j + \frac{1}{2}}{2|I_j|}$$

If we start with a signal that is sampled uniformly over the entire interval $[0, T]$, then n_j will be proportional to $|I_j|$ and each interval will have approximately the same top frequency. However, since the shorter intervals have fewer coefficients, their frequency resolution will be lower. In our experiments, the signal was sampled uniformly at 8 kHz so the maximum detectable frequency was about 4 kHz. The time subintervals ranged in length down to 32 samples, or 4 ms.

A voiced speech signal is produced by regular glottal excitation. Its fundamental frequency is typically in the range 140–250Hz for a female speaker and 100–150Hz for a male speaker [2]. The frequency $F_k = 250$Hz corresponds to about $k = n_j/16$. Using this estimate, we shall introduce a criterion to distinguish voiced from unvoiced speech segments. Define first the frequency index k_0 of the strongest spectral component:

$$|c_{k_0}| = \max_{0 \leq k < n} |c_k|.$$

We will call F_{k_0} the *first fundamental frequency* of the segment.We will say that the signal segment $S_j(t)$ over I_j is *voiced* if $k_0 < n/16$ and *unvoiced* if

$$k_0 > n/16. \tag{7}$$

For more sophisticated recognition problems, we may wish to use more than one frequency to describe the spectrum in a segment. Having found a first fundamental frequency, we suppress all the coefficients at its nearest neighbor frequencies and then look for the strongest survivor. Consider for example a signal $S_j(t)$ over I_j with two fundamental frequencies F_{k_0} and F_{k_1}. To compute F_{k_0} and F_{k_1} we first calculate k_0 from the last equality, set c_k equal to zero if $|k - k_0| < T$ for a preset threshold T (i.e., in a neighborhood of k_0), and then deduce k_1 from

$$|c_{k_1}| = \max_{0 \leq k < n} |c_k|.$$

Finally the two frequencies are obtained via Eq.(6). This procedure may be iterated as long as there are nonzero coefficients c_k. We can force the procedure to terminate earlier by using only the top a per cent of the spectrum, so that relatively small peaks will not show up as fundamental frequencies in any segment. Then T and a are parameters of the algorithm, which can be set empirically, but which should depend upon on the signal-to-noise ratio of the signal. Our experiments typically used $a = 5$ and $T = 5$ or $a = 10$ and $T = 10$.

The coefficients c_k with $|k - k_0| < T$ are associated to the frequency F_{k_0} and contain some extra information. We can average this information into the parameter we extract. Consider the following notion of *center of frequency* associated to each fundamental frequency F_{k_i}:

$$\mu_i = \frac{M_i}{E_i}, \qquad \text{where} \quad E_i = \sum_{k=k_i-T}^{k=k_i+T} c_k^2 \quad \text{and} \quad M_i = \sum_{k=k_i-T}^{k=k_i+T} k c_k^2 \quad \text{for } i = 0, 1, 2, \ldots.$$

In this way, for each fundamental frequency F_{k_i} we can describe its *(locally constant) formants* with the pair $\{\mu_i, E_i\}$ of {center-of-frequency, energy-at-the-frequency}. The *formant representation* for a speech signal consists of a list of intervals together with the top few most energetic locally constant formants for each interval. This data can be used for recognition.

4 Experimental results

We consider the signal corresponding to the first second of the French sentence : *"Des gens se sont levés dans les tribunes"* uttered by a female speaker. The traditional sonogram of this phrase may be obtained as two compressed PostScript files (`sono1.eps.Z` and `sono2.eps.Z`) from the archive site *wuarchive.wustl.edu.* Fig.6 shows the *original signal* on the top part, the *local spectrum* which minimizes entropy in the center and the *reconstructed signal* at the bottom. The reconstructed signal is obtained from the top 5% of the spectrum inside each subinterval. The *adapted time-partition* associated to the local spectrum described in section 2 is drawn with vertical lines.

After finding the adapted time-partition and testing whether the first fundamental frequency in each window is below the cutoff $n_j/16$ of Eq.(7). We then *merge* the adjacent voiced segments and the adjacent unvoiced segments to leave only the windows of Fig.7. This highlights just the transitions between voiced and unvoiced segments.

We note that transitions are found where we expect them. We extracted the time-partitions and listened to the sounds they contain. This gave us the labels in Fig.7, namely /d/ — /e/ — /g/ — /en/ — /s/ — /e/ — /s/. These are alternating voiced and unvoiced segments.

A more sophisticated criterion may be applied to merge adjacent voiced segments which have sufficiently similar formants. This can be done by thresholding with a low-dimensional metric, since the formant representation of each segment has only 4 or 5 parameters in practice. Such a "phoneme recognition" algorithm needs to be well engineered and would probably require some understanding of the speech content to resolve ambiguities, so it is beyond the scope of this communication to discuss it. The present algorithm may be considered as a "front end" to a speech recognition device, intended to simplify the representation of speech down to a few most relevant parameters.

Remark.

On the horizontal axis of Fig.6, we must simultaneously display the sample number, the time

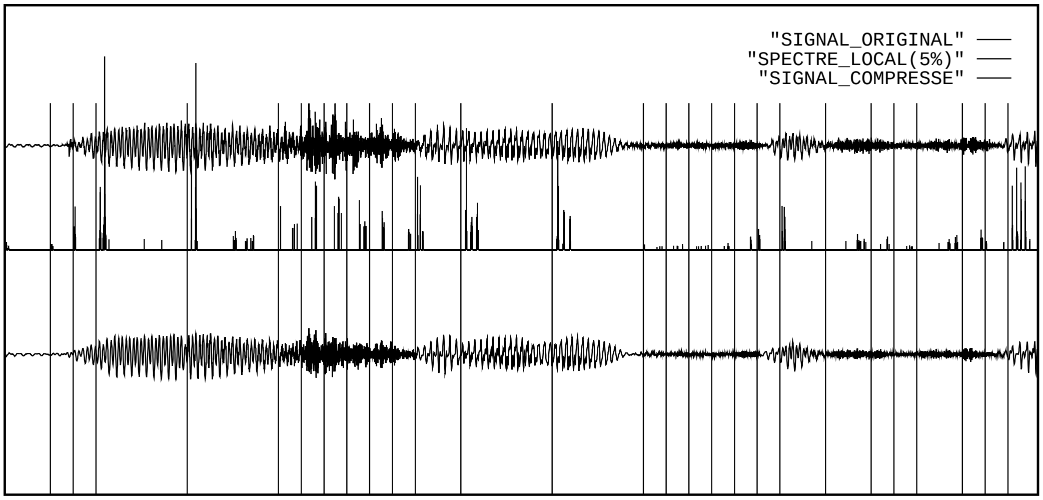

Figure 6: Entropy-minimizing local spectral decomposition

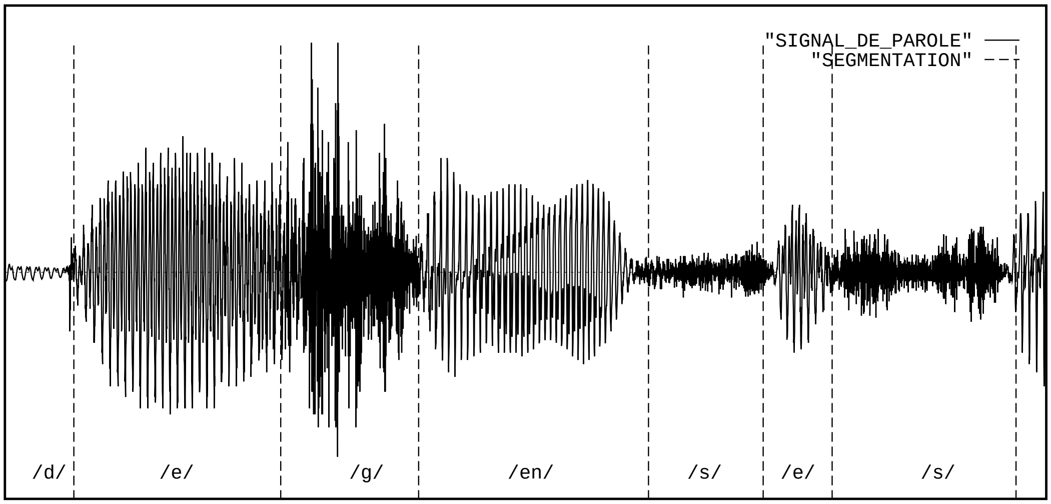

Figure 7: Adapted time-partition after voiced-unvoiced recognition and merging

in milliseconds, the locations of the window boundaries, and the local frequencies within each window. We solve this problem by not using any labels at all. On the topmost and bottommost traces, the horizontal axis represents time. All three traces are intersected by vertical lines indicating the window boundaries, or the endpoints of the intervals I_j in the adapted time-partition. Within each window I_j, position along the horizontal axis of the middle trace of Fig.6 gives the frequency number n_j, which must be scaled by I_j to give the actual frequency. Thus the middle trace is mostly useful for counting the number of formants and gauging their relative strengths and frequencies. This style of presentation is due to Xiang Fang; similar graphs may be seen in [4].

5 Acknowledgments

We are grateful to Yves Meyer for his suggestions and encouragement. One of us (E.W.) would like to express thanks to Christophe D'Alessandro and Nicolas Moreau for helpful discussions and comments during this research, and to Xiang Fang for many discussions during a visit to Washington University in Saint Louis.

List of Figures

References

[1] Pascal Auscher, Guido Weiss, and Mladen Victor Wickerhauser. Local sine and cosine bases of Coifman and Meyer and the construction of smooth wavelets. In Charles K. Chui, editor, *Wavelets–A Tutorial in Theory and Applications*, pages 237–256. Academic Press, Boston, 1992. ISBN 0-12-174590-2.

[2] Calliope. *La parole et son traitement automatique.* Masson, Paris, 1989.

[3] Ronald R. Coifman and Yves Meyer. Remarques sur l'analyse de Fourier à fenêtre. *Comptes Rendus de l'Académie des Sciences*, 312:259–261, 1991.

[4] Ronald R. Coifman and Mladen Victor Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 32:712–718, March 1992.

[5] Christophe D'Alessandro and Xavier Rodet. Synthèse et analyse– synthèse par fonction d'ondes formantiques. *J. Acoustique*, 2:163–169, 1989.

[6] J. S. Liénard. Speech analysis and reconstruction using short time, elementary waveforms. In *Proceedings of IEEE ICASSP-87*, pages 948–951, Dallas, Texas, 1987.

[7] H. Malvar. Lapped transforms for efficient transform/subband coding. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38:969–978, 1990.

[8] K. R. Rao and P. Yip. *Discrete Cosine Transform.* Academic Press, New York, 1990.

[9] Xavier Rodet. Time domain formant-wave-function synthesis. In J. C. Simon, editor, *Spoken Language Generation and Understanding.* D. Reidel Publishing Company, Dordrecht, Holland, 1980.

[10] Mladen Victor Wickerhauser. *INRIA Lectures on Wavelet Packet Algorithms.* INRIA, Roquencourt, France, 1991. Minicourse lecture notes.