Complex Wavelet Based Modulation Analysis

Luneau, Jean-Marc; Lebrun, Jérôme; Jensen, Søren Holdt

Published in:
Proc. of Asilomar Conference on Signals, Systems, and Computers

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Complex Wavelet Based Modulation Analysis

Jean-Marc Luneau†, Jérôme Lebrun‡, and Søren Holdt Jensen†

†*Department of Electronic Systems, Aalborg University, Denmark.*
‡*CNRS - I3S, UMR-6070, Sophia Antipolis, France.*
jml@es.aau.dk

***Abstract*—Low-frequency modulation of sound carry important information for speech and music. The modulation spectrum is commonly obtained by spectral analysis of the sole temporal envelopes of the sub-bands out of a time-frequency analysis. Processing in this domain usually creates undesirable distortions because only the magnitudes are taken into account and the phase data is often neglected. We remedy this problem with the use of a complex wavelet transform as a more appropriate envelope and phase processing tool. Complex wavelets carry both magnitude and phase explicitly with great sparsity and preserve well polynomial trends. Moreover an analytic Hilbert-like transform is possible with complex wavelets implemented as an orthogonal filter bank. By working in an alternative transform domain coined as "Modulation Subbands", this transform shows very promising denoising capabilities and suggests new approaches for joint spectro-temporal analytic processing of slow frequency and phase varying signals.**

I. Introduction

The challenge is to detect, analyze and process slow frequency variations in acoustical cues obtained after a time-frequency analysis. Modulation frequencies of speech between 2 and 16Hz and especially around 4Hz are of great importance for intelligibility [1]. They carry essential syllabic information. On a physiological side, the phase information is also important because of its influence on the human hearing system. The ability to jointly work on these slow modulation frequencies and their phase data is thus crucial for speech. To an other extent, it is also very important for musical acoustic signals [2] or even to such a different field as material surface analysis [3].

The modulation spectrum is commonly obtained in two steps by the spectral analysis of the temporal behavior of the power spectral components. The latter comes first off from a power spectrum analysis (*e.g.*: spectrogram, scalogram, gammatone auditory model). The so-called Complex Modulation Spectrum (CMS) displays time-frequency patterns involving magnitude and phase that reflects different speech articulators or even timbre in music. While the CMS phase information is important for speech intelligibility, its processing is difficult. Modulation filtering requires spectro-temporal tools that jointly work on both CMS- magnitude and phase. In this purpose the Wavelet Modulation Sub-Bands (WMSB) apply a complex wavelet analysis on the temporal trajectories of the time-frequency densities of the signal.

The following paper is organized as follow. First a review of the Complex Modulation Spectrum (CMS) and its physiological importance is done followed by a short introduction to the wavelet theory. Then the proposed method based on Wavelet Modulation Sub-Bands (WMSB) is described with an emphasis on its filtering capacities through an example before concluding with a final discussion.

II. Physiological aspects and Complex Modulation Spectrum

A. Signal phase and cochlea

For acoustical signal processing in general there are important facts to take into account. The first aspect is the signal phase too often ignored when it comes to digital audio processing: two signals with identical magnitude spectra but different phases do sound different. Ohm's acoustic law stating that human hearing is insensitive to phase is persistent but wrong (to a certain extent). For instance, Lindemann and Kates showed [4] that the phase relationships between clusters of sinusoids in a critical band affect its amplitude envelope and most important, affect the firing rate of the inner hair cells (IHC). Thus the major issue is to preserve the phase during an analysis transform otherwise amplitude envelopes will be modified at the reconstruction. Magnitude in a signal gives information about the power while phase is important for localization. For the human hearing, studies like [5] showed that the basilar membrane in the cochlea, basically acts like a weighted map that decomposes, filters and transmits the signal to the IHC. If the phase is altered the mapping on the membrane may be slightly shifted hence the different sounding.

The second important fact for digital audio and speech processing is the mechanical role of the human hearing system and particularly the middle ear and the cochlea. Studies like [6] showed that for frequencies below a threshold of 1.5-2kHz (and gradually up to 6kHz) the firing rate of the IHC depends on the frequency (and on the amplitude and duration) of the stimulus. At those frequencies it is called time-locked activity or phase locking, *i.e.* there is a synchrony between the tone frequency and the auditory nerve response that becomes progressively blurred over this threshold. From 2kHz and above 6kHz, the response of the IHC is function of the stimulus signal envelope and the phase is less important [7].

B. The Complex Modulation Spectrum

The concept of modulation spectrum lies in the spectral analysis of the temporal envelopes of each acoustic frequency

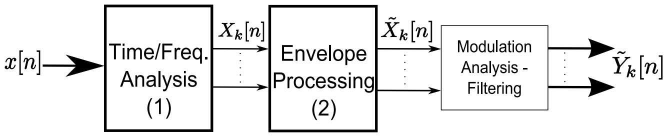

Fig. 1. *Complex Wavelet Based Modulation Analysis*

band. Recent researches have explored three-dimensional energetic signal representations where the second dimension is the frequency and the third is the transform of the time variability of the signal spectrum. The latter is a time-acoustic frequency representation, *i.e.* usually a Fourier decomposition of the signal. The third dimension is the "modulation spectrum", [8] and [9]. The second step of this spectro-temporal decomposition is an envelope processing and can be seen as the spectral analysis of the temporal envelop in each frequency bin. It gives three dimensions to the representation of the signal with two-dimensional energy distributions $S_t(\eta, \omega)$ along time t with η being the modulation frequency and ω the acoustic frequency. Fig. 1 presents a usual modulation analysis approach.

Drullman *et al.* [10], refined later by Greenberg [1], showed that the modulation frequency range of 2-16Hz has an important role in speech intelligibility. It reflects the syllabic temporal structure of speech [1]. More precisely, modulation frequencies around 4Hz seem to be the most important for human speech perception. This is the underlying motivation for effective investigations and further advanced analysis of speech. Those perceptually important spectro-temporal modulations have to be perfectly decorrelated to really open new ways of sparsity for processing as it is showed in the following.

Over the past few years the Complex Modulation Spectrum (CMS) has been a successful tool to analyze important information carried by audio signals unaccessible with usual time-frequency energetic representations. Multiple topics have been investigated with relative success over the last years around modulation frequencies: audio compression [11] , pattern classification and recognition [8], content identification, signal reconstruction, automatic speech recognition, *etc*. In a slightly different nature, modulation frequencies are used to compute the Speech Transmission Index (STI) as a quality measure [12]. It was also experimented in the area of speech enhancement (pre-processing method) to improve the intelligibility in reverberant environments [13] or speech denoising [14] but there again with some limitations. The experiments had to usually face either a production of severe artifacts or a recourse to post-processing because of musical noise.

III. Complex Wavelet Modulation Method

The proposed method is based on a Continuous Wavelet Transform (CoWT) combined with a non-redundant Complex Wavelet Transform (CxWT) (Fig. 2).

A. Wavelet basics

The idea underlying wavelets is to replace the infinitely oscillating sinusoidal basis functions of Fourier transforms by a set of time/scale localized oscillating basis functions obtained by dilatations and translations of a single analysis function, the mother wavelet. A wavelet Ψ is a function of zero average

$$\int_{-\infty}^{+\infty} \Psi(t)dt = 0$$

dilated with a scale parameter s and translated by u

$$\Psi_{u,s}(t) = \frac{1}{\sqrt{s}}\Psi(\frac{t-u}{s}) \tag{1}$$

The wavelet transform of f at the scale s and position u is obtain by convolution of f with the wavelet atom [15]:

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\Psi^*(\frac{u-s}{s})dt \tag{2}$$

A wavelet transform can measure time-frequency variations of spectral components. It may have real or complex coefficients and has a different time-frequency resolution than the windowed Fourier transform. Howevere, the Parseval theorem gives:

$$\begin{aligned} Wf(u,s) &= \int_{-\infty}^{+\infty} f(t)\Psi^*_{u,s}(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{f}(\omega)\hat{\Psi}^*_{u,s}(\omega)d\omega \end{aligned}$$

where $\hat{f}$ is the Fourier transform of f.

In relation to the CMS, in order to analyze the time evolution of frequency tones, it is necessary to use analytic wavelets (*i.e.* whose Fourier transforms are null for negative frequencies) to extract phase and magnitude informations of signal. An analytic function is necessarily complex but is entirely characterized by its real part.

$$f = \Re(f_a)$$

$$\hat{f}_a(\omega) = \frac{\hat{f}_a(\omega) + \hat{f}^*_a(-\omega)}{2}$$

In the time-frequency plane, the energy spread of a wavelet time-frequency atom $\Psi_{u,s}$ is an Heisenberg box of size $s\sigma_t$ along time and σ_ω/s along frequency (σ_t is the time width and σ_ω/s the frequency width). When s varies, the height and width of the rectangle change but the area remains constant ($\sigma_t\sigma_\omega \geq \frac{1}{2}$, uncertainty principle). With these variations it is possible to observe both the amplitudes and their evolutions along time.

This paper focuses on these properties and how to take advantage of them in order to obtain an equivalence of the CMS in the wavelet domain.

B. The Continuous Wavelet Transform

The use of a CoWT at the first step has two roles. It offers a time-frequency density closer to the psychoacoustic model of the human hearing system and provides envelopes with polynomial trends at low and medium frequencies.

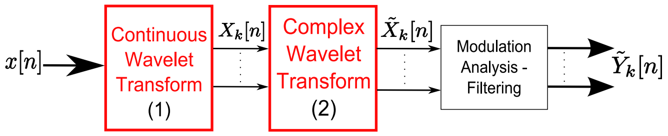

Fig. 2. *Proposed method, CoWT: Continuous Wavelet Transform and CxWT: Complex Wavelet Transform*

1) The CoWT provides a time-scale decomposition of the signal. The log-scale frequency mapping is as such that the low and medium frequencies relevant to speech have a high frequency-resolution and a low time-resolution (Heisenberg uncertainty). It is similar to a localized wide band spectrogram. Meanwhile, the high frequencies less important to speech signals have a better time resolution and lower frequency resolution closer to the human hearing system.
2) With speech and music, formants and harmonics, as amplitudes and tones, evolve slowly along time. This means their envelopes are of polynomial types. In order to capture these slow varying envelopes, the time resolution needs to be low. Thus the low and medium frequencies out of the CoWT have strong polynomial trends.

Not all wavelets are suitable to compute the CoWT. The choice came forward to use the complex Morlet mother wavelet, mostly because it has a bandwidth parameterization. The Morlet wavelet consists of a plane modulated by a gaussian. Equations 3 and 4 give the mother wavelet and its Fourier transform:

$$\Psi_\sigma(t) = C_\sigma \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2}(e^{i\sigma t} - \mathcal{K}_\sigma) \quad (3)$$

$$\hat{\psi}_\sigma(w) = C_\sigma \pi^{-\frac{1}{4}}(e^{-\frac{1}{2}(\sigma-\omega)^2} - \mathcal{K}_\sigma e^{-\frac{1}{2}\omega^2}) \quad (4)$$

with $\sigma = 10$, and $\mathcal{K}_\sigma = e^{-\frac{1}{2}\sigma^2}$ is the admissibility criterion (negligible here).

$$C_\sigma = (1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2})^{\frac{-1}{2}} = 1 \quad (5)$$

is the normalization constant.

The CoWT of a signal $x(t)$ is then defined by:

$$CoWT_\sigma(x) = \int_{-\infty}^{+\infty} \Psi_\sigma(t)x(t)dt =< \Psi_\sigma(t), x(t) > \quad (6)$$

The coefficients obtained from equation 6 would be very redundant if they were not evaluated on a discrete grid of time-scale basis functions. Therefore the CoWT behaves like an orthonormal basis decomposition and it preserves energy. The analyticity and completeness of the CoWT [15] define a local time-frequency energy density which measures the energy of x in the Heisenberg box of each wavelet. This density is called *scalogram*, pendant of the *spectrogram* for the wavelet theory (see Fig. 5)

Furthermore Torrence and Compo [16] showed that synthesis is possible with only the real part of the transform (iCoWT). The reconstructed time signal happens then to be the sum of the real part of the wavelet transform over all scales. As a result, only the magnitude data of the CoWT is preserved.

C. *The Complex Wavelet Transform*

By nature complex wavelets carry both phase and magnitude informations. Furthermore, phase information provides a description of the amplitude and local behavior of a function. Also, an amplitude-phase representation of a function is less oscillatory than the function. And finally because of the important physiological facts seen in II, it is crucial that the second step of the modulation transform provides reliable phase data. The output of the CoWT is thus decomposed using a complex wavelet transform (CxWT) on each scale/frequency bin. The proposed CxWT is implemented via an orthogonal filterbank as shown in Fig. 3. The filterbank has a flexible 3 orthogonal band structure with 2 conjugate high pass filters ($q[n]$ and $q^*[n]$) decimated by 4 to remove the redundancy created by the complex projection [17].

In Fig. 3, $\tilde{X}_{k,0}[n]$ is a coarse version of the sub-band signal $X_k[n]$. The transform distinguishes high- positive and negative, frequencies, $\tilde{X}^+_{k,1}[n]$ and $\tilde{X}^-_{k,1}[n]$ (respectively the positive and negative frequency components of the associated detail signal). They represent Hilbert pairs of wavelets. The complex wavelet filterbank is then iterated N times on each lowpass signal $\tilde{X}_{k,0}[n]$, $\tilde{X}_{k,1}[n]$, $\ldots$ The filterbank creates a complex mapping of the real coefficients from the CoWT.

$h_0[n]$, $h_1[n]$, $g_0[n]$ and $g_1[n]$ are taken to be orthoconjugate complex Daubechies wavelet filters of length 10. They are based on the low-pass filter $g_0[n]$ given in Table I. Furthermore the bandpass orthogonal filter condition on $q[n]$ for analyticity [17] is given by:

$$q[n] := j^n u[n] \quad (7)$$

$$U^*(1/z)U(z) + U^*(-1/z)U(-z) = 2 \quad (8)$$

with

$$u[n] = \frac{\sqrt{3}}{16}[-1, 0, 5, 5, 0, -1] + j\frac{\sqrt{5}}{16}[0, 1, 3, 3, 1, 0] \quad (9)$$

This orthogonal non-redundant CxWT offers a preservation of polynomial trends which is very important after the CoWT as showed in [18]. A good performance on polynomials reflects the good performance of the transform itself [19]. (This behavior on polynomials is directly related to the quality of their projection on a Softy-space, approximation of the Hardy-space.) The CxWT also provides Hilbert transform pairs of wavelets, as well as orthogonality through a realization made of FIR filter approximations to the all-pass IIR filters. As seen in [20], a proper retrieval of the original signal is possible thanks to the perfect reconstruction filterbank, Fig. 4. By its complex nature, the CxWT offers good phase information and improved directionality but no shift invariance. A redundant

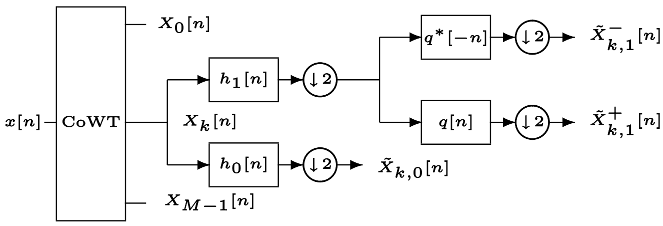

Fig. 3. *Analytic and orthogonal filtebank*

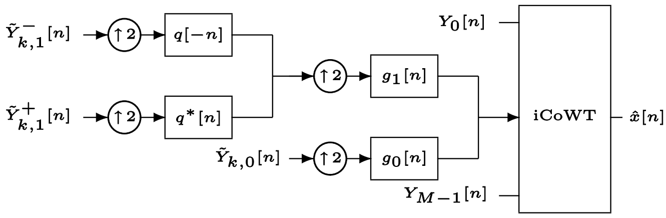

Fig. 4. *Reconstruction filterbank*

higher density implementation is necessary for improved shift invariance [21]. However this implementation of the CxWT could still be a great benefit for applications where sparsity and non-redundancy matter more than shift sensitivity.

TABLE I
COEFFICIENTS FOR ORTHOCONJUGATE COMPLEX DAUBECHIES FILTERS OF LENGTH 10

n	$g_0[n]$	
0	0.01049245051230	+0.02059043708702j
1	-0.00872852869034	-0.01712890812780j
2	0.08063970414533	+0.11794747353812j
3	-0.09422365674476	-0.15137970843150j
4	0.64300323451588	+0.18285216450551j
5	-0.18285216450551	+0.64300323451588j
6	-0.15137970843150	+0.09422365674476j
7	-0.11794747353812	+0.08063970414533j
8	-0.01712890812780	+0.00872852869034j
9	-0.02059043708702	+0.01049245051230j

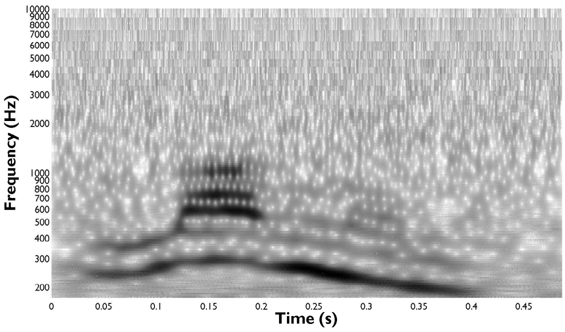

Fig. 5. *Morlet scalogram of the word “longing” with noise*

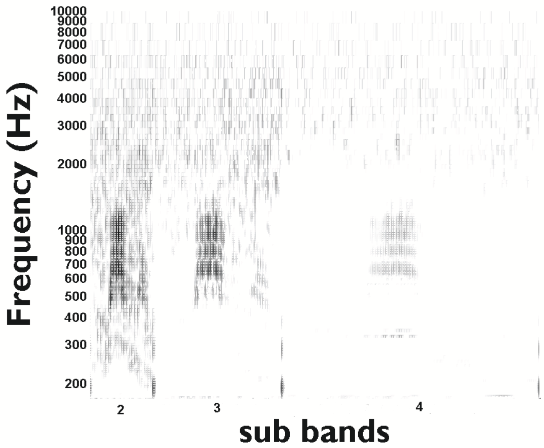

Fig. 6. *Modulation subbands with scalogram replicas*

IV. Results and Capabilities

A. Representation

To illustrate the modulation transform nature of the proposed method, a recording of the word “longing” drowned in white noise and sampled at 22050Hz has been analyzed. Fig. 5 shows the first step of the transform, *i.e.* the scalogram resulting from the Morlet CoWT. Only the magnitude is displayed as the phase is not indispensable at that stage. As explained previously, the phase data is only important in the second step of the transform.

Fig. 6 illustrates the modulation subbands. The continuous DC part and the very low modulation frequencies, in the first subband, are not shown as they represent too much energy in comparison. The important observation relates to the sparsity of the decomposition and how replicas of the scalogram appear. In each subband a replica shows the corresponding modulation frequency range/scale that is in the scalogram. Fig. 6 only shows the magnitude of the coefficients but each coefficient is a complex pair carrying explicitly both magnitude and phase data.

B. Processing capabilities

This representation (Fig. 6) offers different possibilities of processing: estimation, detection, denoising, compression, enhancement by energy growth *etc.* Each subband may also be processed independently. Every scale and replica are independent thanks to the orthogonality of the decomposition. Low modulation frequencies are important for intelligibility while higher ones (100-200Hz) show the fundamental frequency of the talker.

This paper focuses on low modulation frequencies due to the log-scale decomposition due to the wavelet multi-resolution behavior. Lower modulation scales represent a shorter modulation-frequency range. With Daubechies filters of length 10, the 2-3 first subbands are important to speech (modulation frequency range of 2-12Hz). Shorter filters would give more precision in the low modulation frequencies but the complex mapping projection would not be as good. Hence, only filter lengths of 10 are used as a good compromise between precision and preservation of polynomial trends.

C. Denoising by thresholding

Two different estimations can be made on the complex modulation subbands depending on the aim. Hard or soft thresholding should be used based on the local energy density of the wavelet coefficients.

Hard thresholding

$$\tilde{Y}_k[n] = \begin{cases} \tilde{X}_k[n] & \text{if } |\tilde{X}_k[n]| > T \\ 0 & \text{if } |\tilde{X}_k[n]| \leq T \end{cases} \tag{10}$$

Soft thresholding

$$\tilde{Y}_k[n] = \begin{cases} \tilde{X}_k[n] - T & \text{if } |\tilde{X}_k[n]| > T \\ 0 & \text{if } |\tilde{X}_k[n]| \leq T \end{cases} \tag{11}$$

where T is of the form $\sigma\sqrt{2\log_e N}$ (with σ^2 the variance of $\tilde{X}_k$ and N the size of the reconstruction basis [15]). Hard thresholding is used when only the energetic coefficients inside the modulation subbands need to be preserved. Hard thresholding as well as a removal of whole subbands is used for strong noise or high data reduction. For softer noise or lower data reduction it is preferable to use soft thresholding. In that case indeed, T will be chosen with the highest probability to be above the low coefficients. So that they are considered to be noise-like and the thresholding will have a denoising effect [18] on the scale/frequency bins. Different thresholds can also be applied depending on the targeted acoustical frequency range. Fig. 6 shows that most of the noise is decomposed in the high frequencies that are less important to intelligibility. So high frequencies can usually undergo a heavier processing.

As the thresholding jointly works on both magnitude and phase data, disturbing artifacts or musical noise are avoided. Even keeping only the first subband (DC and very low modulation frequencies) yields a speech signal of poor quality but still intelligible. This confirms Greenberg's [1] and Steeneken's [12] works on the role of low modulation frequencies (4Hz) in speech. .Naturally most advanced wavelet tools for denoising and estimation may also be used in a data reduction goal. More advanced schemes would also provide scalability features in the estimation/quality process.

V. Conclusion and Discussion

This paper presented a complex valued transform for speech compression based on modulation frequencies. Low modulation frequencies contain crucial cues for speech intelligibility so the idea was to exploit that property in combination to the great sparsity of the non redundant complex wavelet transform. It can primarily be used for speech denoising in a modulation subband approach but also shows interesting capabilities for compression. The CxWT offers useful phase information to the Complex Modulation "Spectrum" that allows joint work on both magnitude and phase. Precisely what was needed to do filtering in the modulation domain. It still does not have shift invariance but provides a preservation of polynomial trends, Hilbert-like pairs of coefficients, orthogonality and uses FIR filters. However the Hilbert pairs are not perfect and show little aliasing energy in the negative frequency range [17] that might affect the reliability of the amplitude and phase information.

Nevertheless, the method suggests an alternative approach modulation filtering of speech and slow varying audio signals. The joint magnitude/phase processing overcomes the distortions encountered by the usual approaches and promises efficient means for modulation denoising.

Acknowledgment

The work of Jean-Marc Luneau is supported by the EU by a Marie-Curie Fellowship (EST-SIGNAL program : http://est-signal.i3s.unice.fr) under contract No MEST-CT-2005-021175.

References

[1] S. Greenberg, "On the origins of speech intelligibility in the real world," *ESCA Workshop on robust speech recognition for unknown communication channels*, pp. 22–32, 1997.

[2] L. Navarro, G. Courbebaisse, and J.-C. Pinoli, "Continuous frequency and phase spectrograms: a study of their 2D and 3D capabilities and application to musical signal analysis," *Journal of Zhejiang Univ. Sci. A*, vol. 9(2), pp. 199–206, Jan 2008.

[3] F. Leonard, "Spectrogramme de phase et spectrogramme de frequence," *Revue du Traitement du Signal*, vol. 17(4), pp. 269–286, 2001.

[4] E. Lindemann and J. M. Kates, "Phase relationships and amplitude envelopes in auditory perception," *WASPAA*, pp. 223–226, 1999.

[5] L. Golipour and S. Gazor, "A biophysical model of the human cochlea for speech stimulus using stft," *ISSPIT*, 2005.

[6] J. Blauert, *Spatial hearing: The psychophysics of human sound localization*, M. Press, Ed. MIT Press, 1996.

[7] D. H. Johnson, "The relationship between spike rate and synchrony of auditory-nerve fibers to single tones," *JASA*, vol. 68 (4), pp. 1115 – 1122, 1980.

[8] S. Sukkittanon, L. E. Atlas, and J. Pitton, "Modulation-scale analysis for content identification," *IEEE Trans. Sig. Proc.*, 2004.

[9] H. Hermansky, "The modulation spectrum in the automatic recognition of speech," *Automatic Speech Recognition and Understanding*, vol. 1, pp. 140–147, 1997.

[10] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *JASA*, vol. 95, pp. 1053–1054, 1994.

[11] M. Vinton and L. Atlas, "A scalable and progressive audio codec," *ICASSP*, vol. 5, pp. 3277–3280, 2001.

[12] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," *JASA*, vol. 67, pp. 318–326, 1980.

[13] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *Speech Communication*, vol. 45 (2), 2005.

[14] H. Hermansky, E. A. Wan, and C. Avendano, "Speech enhancement based on temporal processing," *ICASSP*, pp. 405–408, 1995.

[15] S. Mallat, *Une exploration des signaux en ondelettes*, 2nd ed. Ecole Polytechnique, 1999.

[16] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *BUlletin of the American Meteorological Society*, vol. 79, pp. 61–78, 1998.

[17] R. van Spaendonck, T. Blu, R. Baraniuk, and M. Vetterli, "Orthogonal Hilbert transform filter banks and wavelets," *ICASSP*, vol. 6, pp. 505–508, Apr 2003.

[18] J.-M. Luneau, J. Lebrun, and S. H. Jensen, "Complex wavelet modulation sub-bands and speech," *ISCA ITRW on Speech Analysis and Processing for Knowledge Discovery, Aalborg*, vol. 1, 2008.

[19] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[20] J.-M. Luneau, J. Lebrun, and S. H. Jensen, "Complex wavelet based envelope analysis for analytic spectro-temporal signal processing," *Technical Report, Aalborg University*, vol. R08-1001, ISSN: 0908-1224, 2008.

[21] I. W. Selesnick, "A higher-density discrete wavelet transform," *IEEE Trans. on Signal Processing*, vol. 54 (8), pp. 3039–3048, 2006.