# Complex Wavelet Modulation Sub-Bands and Speech

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

*Published in:*
Proceedings for ISCA ITRW Speech Analysis and Processing for Knowledge Discovery

*Publication date:*
2008

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

# Complex Wavelet Modulation Sub-Bands and Speech

*Jean-Marc Luneau*[1]*, Jérôme Lebrun*[2] *and Søren Holdt Jensen*[1]

[1]Department of Electronic Systems, Aalborg University, DK-9220 Aalborg, Denmark.
[2]UMR-6070, CNRS, FR-06903 Sophia Antipolis, France.
jml@es.aau.dk

## Abstract

A new class of signal transforms called Modulation Transforms has recently been introduced. They add a new dimension to the classical time/frequency representations, namely the modulation spectrum. Although very efficient to deal with different applications like feature extraction, speech recognition and also analysis for audio coding, these transforms show their limits *e.g.* when used to remove non-trivial noise from speech signals. Modulation sub-band decompositions based on the computation of the Hilbert envelope have been proved to create disturbing artifacts. We detail a new method to deal properly with the phase and the magnitude of the modulation spectrum in a linear and analytic framework based on a complex wavelet transform. This Complex Wavelet Modulation Sub-Band transform gives some interesting results in speech denoising and proposes a new approach for analytic signal processing in general.
**Index Terms**: speech analysis, complex wavelets, modulation spectrum, denoising, phase signal.

## 1. Introduction

Much of speech processing has been relying on the use of spectral analysis. However efficient in the frequency domain, this approach shows some limitations when it comes to provide a deeper understanding of the whole perception/production mechanisms for sound and speech signals. In classical spectral analysis, many important temporal properties of these signals are structurally occulted and this is the reason why some years ago, after the work of Steeneken and Houtgast on the Speech Transmission Index (STI) [1], investigations started in the direction of joint spectro-temporal analysis [2, 3]. The goal was, and is still, to determine the interactions between temporal and acoustic frequencies cues.

Many natural signals can be seen as the sum of low frequency modulators of higher frequency carriers. This concept of modulation frequency appears to be very useful to represent and analyze broadband acoustic signals. The starting point in this paper is to understand where the focus should be put on during the so-called modulation frequency analysis of speech to make it reliable and efficient.

Important physiological facts together with an introduction to modulation frequencies will first be presented. The need for other ways to build the transform inspired by the principle of coherent detection [4] will then be stressed. To get analyticity, we will explain the motive of using complex wavelets instead of real-valued ones or classical Fourier analysis. Then we will show some interesting outcomes of this Complex Wavelet Modulation Sub-Band transform and conclude with its legitimacy in forthcoming applications for speech and general signal processing.

* This work was supported by the EU via a Marie Curie Fellowship (EST-SIGNAL program http://est-signal.i3s.unice.fr) under contract No MEST-CT-2005-021175.)

## 2. Physiological background and approach

### 2.1. 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. For instance, Lindemann and Kates showed in 1999 [5] 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 a modulation transform otherwise amplitude envelopes will be modified. Magnitude in a signal gives information about the power while phase is important for localization. For the human hearing, studies like [6] 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. Different studies [7] showed that for frequencies below approximately 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 [8].

### 2.2. Modulation frequencies

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" [9, 10] . The second step of this spectro-temporal decomposition can be viewed as the spectral analysis of the temporal envelop in each frequency bin. It gives a three-dimensional representation of the signal with two-dimensional energy distributions $S_t(\eta, \omega)$ along time $t$ with $\eta$ being the modulation frequency and $\omega$ the

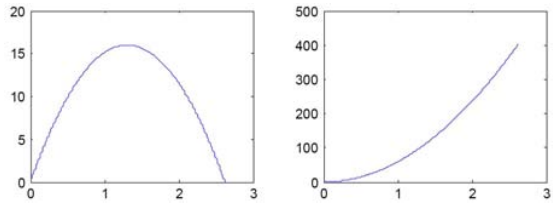



Figure 1: *Amplitude and phase of a complex chirp model for voiced speech signals resulting from the first transform.*

acoustic frequency.

Drullman *et al.* [11], refined later by Greenberg [3], showed that the modulation frequency range of 2-16Hz has an important role in speech intelligibility. It reflects the syllabic temporal structure of speech [3]. More precisely, modulation frequencies around 4 Hz seem to be the most important for human speech perception. Low frequency modulations of sound seem to carry significant information for speech and music. This is the underlying motivation for effective investigations and further advanced analysis in speech enhancement. Those perceptually important spectro-temporal modulations have to be perfectly decorrelated to really open new ways for processing as we show it in the following.

Multiple topics have been investigated with relative success over the last years with this transform: pattern classification and recognition [9], content identification, signal reconstruction, audio compression, automatic speech recognition *etc*. In a slightly different manner, modulation frequencies are used in order to compute the Speech Transmission Index (STI) as a quality measure [1]. It was also experimented in the area of speech enhancement (pre-processing method) to improve the intelligibility in reverberant environments [12] or speech denoising [13] 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.

## 3. Analyticity and cohesion

### 3.1. Limitations of the standard Hilbert envelope approach

The usual modulation spectrum frameworks relies on envelope detection based on the analytic signal or quadrature representation obtained from the Hilbert envelope in each sub-band. Theses approaches are easy to interpret but their modulators are always non-negative and real. More precisely, the input signal $x$ is decomposed into $M$ sub-band signals $x_k$ using typically a bank of modulated filters $h_k$ where $k = 0, \ldots, M-1$ is the sub-band index. When using real filters, the extraction of the envelope in each sub-band is done via the Hilbert transform $H\{.\}$ by introducing $\tilde{x}_k := x_k + jH\{x_k\}$, *i.e.* the analytical extension of $x_k$. Now, each sub-band signal can be decomposed into its envelope $m_k := |\tilde{x}_k|$ and its instantaneous phase $p_k := \cos(\varphi_k)$ with $\tilde{x}_k = m_k \exp(j\varphi_k)$. We get $x_k = m_k p_k$.

The modulation spectrum in the $k^{th}$ sub-band is then the spectrum of the envelope signal $m_k$. With this approach, any filtering or processing of the sub-bands introduces artifacts and distortions at the reconstruction. As stressed in [4], this is essentially due to the way the envelope signal is obtained. Processing the modulation spectra without taking great care of the phase signals $p_k$ leads to a leakage of energy from the modified sub-band onto the others. The reconstructed sub-band being the product in time-domain of the modified envelope and the original carrier (giving a convolution in the Fourier domain), the bandwidth of the modified sub-band may be widened. This leads to imperfect alias cancellation between the sub-bands and thus artifacts.

Schimmel and Atlas [4] proposed to reconstruct narrow-bandwidth sub-bands to achieve little leakage. They suggested the use of a "coherent" carrier detection to get a $\tilde{\varphi}_k$ close to the true phase of the signal but also narrow-band. Thus, both the envelope and the carrier must be complex, so the envelope and phase seen previously become $m_k^c := \tilde{x}_k \exp(-j\tilde{\varphi}_k)$ and $p_k^c(t) = \exp(j\tilde{\varphi}_k)$ where $\tilde{\varphi}_k$ is a low-pass filtered version of the estimated phase signal. Their idea is to design this low-pass filter by compromising the desired amount of distortion and the effectiveness of modulation filters stop-band attenuation.

### 3.2. Necessity of a new approach

We introduce here an alternative method that completely avoids the issue of computing the envelope signals but nevertheless provides a time-scale version of the modulation spectrum for each sub-band. The underlying idea in our approach is motivated by the fact that for speech/voiced signals, extracting the envelopes of the sub-band signals, is similar to extracting their polynomial parts. Namely, the sub-band signals out of the first transform resemble $c(t) = w(t).e^{j2\pi(\omega_1 t^r + \omega_0)}$ (Fig. 1) where $w(t)$ is a piecewise polynomial envelope, $\omega_0, \omega_1$ frequency parameters and $r$ characterizes the frequency evolution. This gives a good model for the sub-bands signals for voiced speech. Hence, the principle will be to perform complex-wavelet transforms on each band to extract the polynomial part while dealing properly with the phase as we detailed it in [14]. The difference here is that the spectro-temporal approach proposed cannot be called modulation spectrum as we work with polynomials approximations coming from wavelet processing. There is no actual spectrum after the second part of the transform. With this new approach, improvements should be possible in many spectro-temporal related application domains and not only in speech enhancement.

## 4. Complex wavelet method

The problem with most spectro-temporal or modulation frequency frameworks is often the lack of resolution at the crucial low modulation frequencies. This drawback comes again from using the Fourier analysis as second transform in the process as it only permits a uniform frequency decomposition which yields to uniform modulation frequency resolution. A log frequency scale allows to adapt the precision on the important modulation frequencies between 2 and 16 Hz [9]. Moreover, from a psychoacoustic point of view [15], such a scale best matches the human perceptual model of modulation frequencies, hence again the idea of using a wavelet transform as second step of the modulation transform, especially for natural or speech signals.

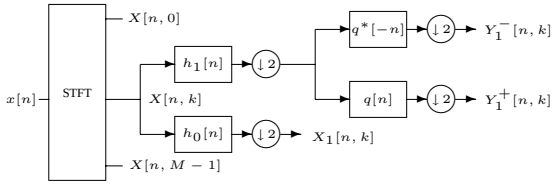
The discrete wavelet transform has been a successful new tool in many fields of signal processing and especially in image processing. In brief, the idea underlying wavelets is to replace the infinitely oscillating sinusoidal basis functions of Fourier-like transforms by a set of time/scale localized oscillating basis functions obtained by the dilatations and translations of a single analysis function, the *wavelet*. Nevertheless, with the first generation of real-valued wavelets, it was difficult to deal properly with both amplitude and phase informations in a signal. This explains partly their limited success in audio and speech pro-

cessing. However, the recent developments of new complex-valued wavelet-based transforms [16] alleviates most of these limitations. Complex wavelets have the property to deal properly with both amplitude and phase of the signal which is a crucial matter as seen earlier.

It has been shown that by using complex wavelets, one can implement new filterbank structures that ensure the analyticity of the analysis [17]. As usual, the filterbank will be used in an iterated manner [14]. Indeed, the analysis of a signal at several scales (multi-resolution analysis) consists of iterating the filterbank on the low-pass sub-band and cascading it up to a certain level $l$ of details.

## 5. Speech denoising experiment

Here, by working with complex wavelets, we avoid the limitations of the usual Hilbert envelope approaches caused by the separate processing on the magnitude and the phase of the modulation spectrum in the sub-bands. In our approach the sub-band signals $X_k[n] = X[n,k]$ are obtained using a complex modulated filter-bank (a Short Time Fourier Transform here) for $k = 0, \ldots, M-1$ and further decomposed using an orthogonal complex wavelet filterbank as shown in




where $X_1[n,k]$ is a coarse version of the sub-band signal $X[n,k]$ and $Y_1^+[n,k]$ and $Y_1^-[n,k]$ are respectively the positive and negative frequency components of the associated detail signal. The complex wavelet filterbank is then iterated $N$ times on each lowpass signal obtained $X_1[n,k], X_2[n,k], \ldots$. Here, motivated by their good phase behavior, we took $h_0[n]$, $h_1[n]$, $g_0[n]$ and $g_1[n]$ to be orthoconjugate complex Daubechies wavelet filters. More precisely, we did our experiments using the complex Daubechies filters of length 10 based on the the low-pass filter $g_0[n]$, see [14].

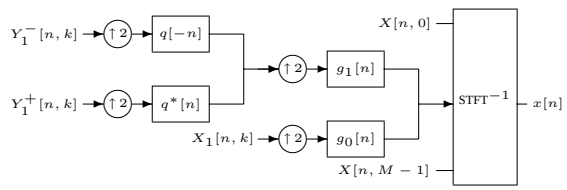
Now, $q[n]$ is a bandpass orthogonal filter that satisfies the conditions given in [17] to get analyticity, *i.e.* it is obtained from a complex-valued lowpass orthogonal filter $u[n]$ satisfying

$$U^*(1/z)U(z) + U^*(-1/z)U(-z) = 2.$$

In our case we took $q[n] := j^n u[n]$ where

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

The reconstruction is then done using the complementary synthesis filterbank.




Now, we omit some details of the signals at the reconstruction by picking only the *relevant* coefficients in the decomposition - this is the underlying principle of denoising by sparse representations. Indeed, for a *well designed* reconstruction basis, the noise is not be picked in the sparse coefficients used to reconstruct the signal, hence the denoising. The "quality" of the reconstructed signal depends largely on the choice of the basis vectors with which the reconstruction is performed. In our case, the dual stage synthesis, inverse Complex-DWT followed by inverse STFT, gives reconstruction vectors that are well adapted to acoustical signal processing, namely dilated windowed sinusoidal functions similar to scaled Gabor functions.

Furthermore, this decomposition separates the complex-valued components obtained (*i.e.* with proper magnitude and phase) into orthogonal spaces. With this method, if we do any thresholding or remove sub-bands from the wavelet decomposition, we do not create aliasing problems between the sub-bands. Typically, if some uncorrelated noise is spread on the modulation sub-bands, for each of them the phase and the magnitude can be properly cleaned [14]. We are thus insured not to widen the spectral bandwidth of the sub-band and thus not to smear on the near-close sub-bands.

### 5.1. Wavelet thresholding

It is now possible to work on the coefficients using all the wavelet related tools for denoising and especially thresholding, in a hard or a soft manner. For example here, from Fig.2 to Fig.3, to evaluate the denoising capacities of the framework, a simple hard thresholding has been applied on every sub-band. The hard threshold used is of the form $T = \sigma\sqrt{2\log_e N}$ (with $\sigma^2$ the noise variance and $N$ the size of the basis we reconstruct with, [18])

Sound files available at this URL: www.luneau.info/ITRW and the presented spectrograms seemed to be the best way to show the results of the denoising in the absence of formal listening test. They differ from the usual masking approach because we can work effectively on all the resolutions of the modulation spectra without smearing on the other ones. Between the two spectrograms we can see that some structures in the high frequencies have been removed. The very high coefficients have been attenuated but the visible structures of the speech signal are still present. Because the denoising was basic and the original speech signal drowned in important urban noise, only little information above 3.5-4kHz remains and the sound quality obtained is more intelligible but can be improved.

## 6. Discussion

Neither formal listening tests nor computation of the Speech Transmission Index (STI) [1] have been performed so far. Our scope was to build a processing framework rather than focusing on intelligibility. The denoising methods have to be fine-tuned before setting up listening tests. The results of this thresholding in the modulation domain are very encouraging nonetheless. The urban noise that alters the signal is one of the most difficult to get rid of, but with this technique, a big part of it was removed without production of annoying artifacts. Only informal listening tests have been performed. We lost a bit of the natural sounding of the speech signal because we did not take enough care of the important low modulation frequencies. This will be improved in the future. An other issue is the representation of the transform. By nature, the three dimensions of the modulation sub-bands with the multi-scale complex wavelet decom-

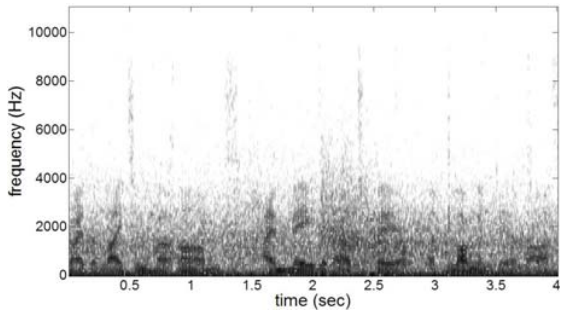



Figure 2: *Spectrogram of the noisy signal*

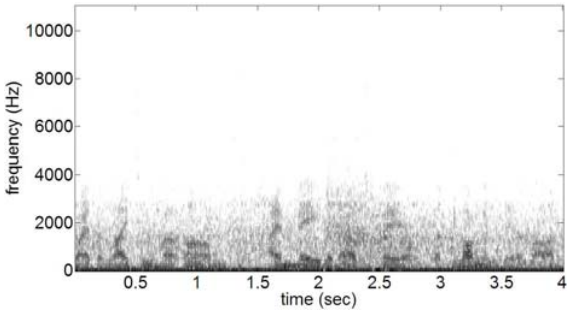



Figure 3: *Spectrogram of the denoised signal*

position are problematic to represent. Furthermore, the process starts with a time-frequency decomposition, a STFT, which may not be optimal for speech data. Traditionally speech signals are processed with 20msec segments. But the important modulation frequencies are between 2 and 12Hz, which means durations between 8 and 50msec. This means that not only the second part of the transform is crucial but the first time-frequency decomposition is also very significant. Hence, in a near future, focus will be put on making the first transform (STFT) more compatible with the complex wavelet transform in terms of magnitude and phase information acquired from it.

## 7. Conclusion

In this paper we introduced a new way to process modulation frequencies using complex wavelets. We proved the legitimacy of this approach since the transform we proposed is based on complex wavelets which deal properly with phase and magnitude informations in the signal. This new signal representation gives a linear access to the three dimensions of the transform at the same time. The proposed framework was also tested on some speech signal drowned in urban noise and the results illustrate the new possibilities for speech denoising but also compression and probably many more topics. So far we have defined a rather general and simple framework for our first experiments but in the near future the transform and the resulting representation will be improved in order to enable the use of more sophisticated denoising tools coming from the wavelet theory and used especially in image processing.

## 8. References

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

[2] T. Houtgast and H. J. M Steeneken, "A review of the mtf concept in room acoustics," *JASA*, vol. 77, pp. 1069–77, 1985.

[3] 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.

[4] S. Schimmel and L. E. Atlas, "Coherent envelope detection for modulation filtering of speech," *ICASSP*, vol. 1, pp. 221–224, mar 2005.

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

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

[7] J. Blauert, "Spatial hearing: The psychophysics of human sound localization," *MIT Press*, 1997.

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

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

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

[11] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech perception," *JASA*, vol. 95, pp. 1053–64, February 1994.

[12] 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), Feb 2005.

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

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

[15] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *JASA*, vol. 85, pp. 1676–80, 1989.

[16] I. W. Selesnick, R. G. Baraniuk, and N. Kingsbury, "The dual-tree complex wavelet transform - a coherent framework for multiscale signal and image processing," *IEEE Signal Processing Magazine*, vol. 22(6), pp. 123–151, Nov 2005.

[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] S. Mallat, "A wavelet tour of signal processing," *Academic Press*, 1999.