Complex Wavelet Based Modulation Analysis of Speech

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

[1] *Department of Electronic Systems, Aalborg University, Denmark.*

[2] *CNRS - I3S, UMR-6070, Sophia Antipolis, France.*

jml@es.aau.dk

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 and musical acoustic signals.

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 timbre in music. While the CMS-envelope 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) method 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) 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.

Complex Modulation Spectrum

The concept of modulation spectrum lies in the spectral analysis of the temporal envelopes of each acoustic frequency 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" [2]. The second step of this spectro-temporal decomposition is an envelope processing and can be seen as the spectral

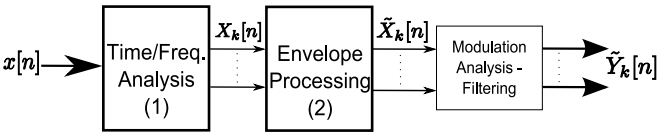

Figure 1: *Complex Wavelet Based Modulation Analysis*

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. Figure 1 presents a usual modulation analysis approach.

Drullman *et al.* [3], 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 CMS has been used to analyze important information carried by audio signals unaccessible with usual time-frequency energetic representations. Relatively successful investigations over the last years around modulation frequencies include: audio compression [4] , pattern classification and recognition [2], content identification, signal reconstruction, automatic speech recognition, *etc.* In a slightly different nature, modulation frequencies help to compute the Speech Transmission Index (STI) as a quality measure [5]. It was also experimented in the area of speech enhancement (pre-processing method) to improve the intelligibility in reverberant environments or speech denoising [6] but there again with some artifact limitations.

Complex Wavelet Method

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.

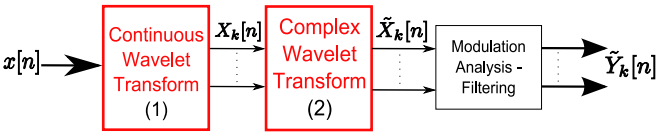

Figure 2: *Proposed method, CoWT: Continuous Wavelet Transform and CxWT: Complex Wavelet Transform*

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.

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

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 smooth polynomial trends at low and medium frequencies.

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 smooth polynomials. 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.

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 1 and 2 give the mother wavelet and its Fourier transform:

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

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

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

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

is the normalization constant.

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

$$CoWT_\sigma(x) = \int_{-\infty}^{+\infty} \Psi_\sigma(t)x(t)dt = < \Psi_\sigma(t), x(t) > \quad (4)$$

The coefficients obtained from equation 4 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 [7] 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 Figure 4)

Furthermore Torrence and Compo [8] 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.

The Complex Wavelet Filterbank

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 important physiological facts , 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 Figure 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 [9].

In Figure 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 $\tilde{X}_{k,0}[n]$, $\tilde{X}_{k,1}[n]$, ... 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. Furthermore the bandpass orthogonal filter condition on $q[n]$ for analyticity [9] is given by:

$$q[n] := j^n u[n] \quad (5)$$

$$U^*(1/z)U(z) + U^*(-1/z)U(-z) = 2 \quad (6)$$

with

$$u[n] = \frac{\sqrt{3}}{16}[-1, 0, 5, 5, 0, -1] + j\frac{\sqrt{5}}{16}[0, 1, 3, 3, 1, 0] \quad (7)$$

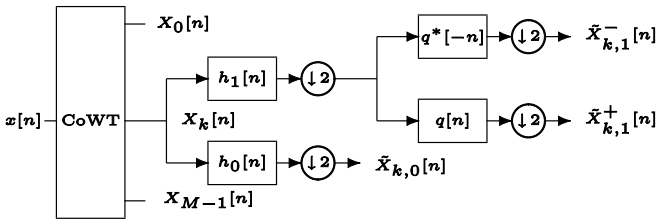

Figure 3: *Analytic and orthogonal filterbank*

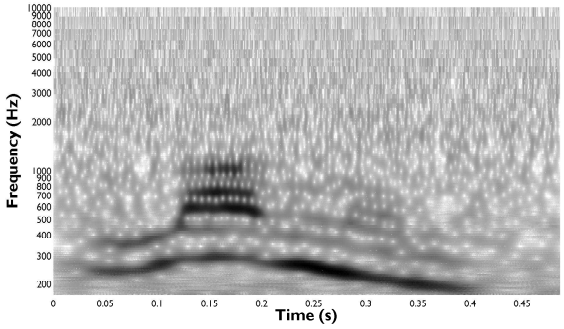

Figure 4: *Morlet scalogram of the word "longing" with noise*

This orthogonal non-redundant CxWT offers a preservation of polynomial trends which is very important after the CoWT as showed in [10]. A good performance on polynomials reflects the good performance of the transform itself. 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 [11], a proper retrieval of the original signal is possible thanks to a perfect reconstruction filterbank. By its complex nature, the CxWT offers good phase information and improved directionality but no shift invariance. A redundant higher density implementation is necessary for improved shift invariance. However this implementation of the CxWT could still be a great benefit for applications where sparsity and non-redundancy matter more than shift sensitivity.

Envelope Detection

It is important to note that an actual modulation spectrum analysis requires envelope detection. Three possible ways exist to do so:

Incoherent : magnitude of the Hilbert envelope, not band-limited.

Coherent : frequency shift of the complex spectrum down to the DC.

Decimation : frequency shift via decimation.

The Complex Wavelet Method, within the filterbank and thanks to the so called noble identities takes profit of the decimation steps as an envelope detection.

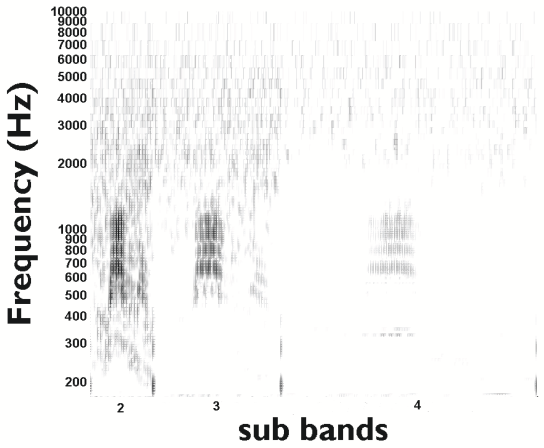

Figure 5: *Modulation subbands with scalogram replicas*

Results and Capabilities

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. Figure 4 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.

Figure 5 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. Figure 5 only shows the magnitude of the coefficients but each coefficient is a complex pair carrying explicitly both magnitude and phase data.

Processing capabilities

This representation (Figure 5) 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 from 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.

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

$$\tilde{Y}_k[n] = \begin{cases} \tilde{X}_k[n] & \text{if } |\tilde{X}_k[n]| > T \\ 0 & \text{if } |\tilde{X}_k[n]| \leq T \end{cases} \quad (8)$$

Soft thresholding

$$\tilde{Y}_k[n] = \begin{cases} \tilde{X}_k[n] - T & \text{if } |\tilde{X}_k[n]| > T \\ 0 & \text{if } |\tilde{X}_k[n]| \leq T \end{cases} \quad (9)$$

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 [7]). 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 [10] on the scale/frequency bins. Different thresholds can also be applied depending on the targeted acoustical frequency range. Figure 5 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 [5] 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.

Conclusion and discussion

This paper presented a complex valued transform for speech analysis 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 [9] 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 processing.

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*, pp. 22–32, 1997.

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

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

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

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

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

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

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

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

[10] 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.

[11] ——, "Complex wavelet based envelope analysis for analytic spectro-temporal signal processing," *Technical Report, Aalborg University*, vol. R08-1001, ISSN: 0908-1224, 2008.