Coherent Modulation Filtering For Speech

Qin Li and *Les Atlas*
Department of Electrical Engineering, University of Washington
Box 352500, Seattle, WA 98195-2500

ABSTRACT

Modulation filtering ideally offers a new approach to modifying the dynamics of non-stationary signals, such as speech. In this paper, a new type of coherent modulation analysis and filtering method is proposed. The new method consists of two essential parts – an instantaneous frequency estimator based on conditional mean frequency, which is used for coherent modulation analysis, and a multi-component decomposition based on spectrogram peak tracking, which is used to separate multiple modulation components in signals. An important modulation filtering property, frequency shift invariance, is achieved with the new proposed method.

Index Terms **-** Speech analysis, Modulation, Time-varying filters, Time-frequency analysis.

1. INTRODUCTION

Many natural occurring and man-made signals can be represented by a modulation model, where low-frequency signals modulate (multiply) high-frequency carriers. The concepts of "modulation frequency" and "modulation filtering" are associated with Fourier analysis and linear filtering of the low frequency modulators, respectively. Modulation frequency analysis and filtering are useful and potentially powerful new tools for describing, representing and modifying broadband acoustic signals.

One critical step of modulation analysis and filtering techniques is to separate broadband signals into modulators and carriers. A commonly used method is Hilbert envelope detection, in which broadband signals are first decomposed into frequency subbands and each subband's modulator signal is assumed real and positive. However, recent studies have shown that in many signals this real-and-positive-envelope assumption is inaccurate and consequently induces much phase discontinuity and audible distortion results from methods which modify these envelopes [1, 2]. Also as reported by Ghitza [3] and Schimmel *et al* [4], the Hilbert envelop based methods show considerably less modulation frequency attenuation than intended.

Lately, a new concept of coherent modulation analysis and filtering has been introduced [5], and several coherent methods have been proposed and tested [4-6]. The distinguishing aspect of these coherent methods is that the modulator is no longer assumed to be real and positive and is instead, in general, complex. These new approaches offer advantages over previous Hilbert envelope based methods.

This work was partially supported by Air Force Office of Scientific Research Grant FA95500610191.

It remains a key research challenge to design a good coherent demodulation method. With no other constraints, there are an infinite number of ways to decompose a signal into a modulator and a carrier. This carrier-modulator separation problem is closely related to instantaneous frequency (IF) estimation [4, 6] and the recent IF-based results have offered improved results and less processing distortion.

One important aspect that has not received much quantitative attention in modulation filtering techniques is multi-component decomposition. Most naturally occurring acoustic signals, e.g. speech signals, are broadband signals usually containing multiple modulation components (e.g. multiple harmonics). In previous studies, fixed filter banks were commonly used to decompose the signal into subbands and perform modulation analysis and filtering within each subband. However, it is very common that carrier signals can cross the frequency subband boundaries. When a single component crosses two or more subbands and the modulation filtering is done separately in each subband, significant discontinuities can appear at the places where the carrier frequencies cross the subband boundaries.

In this paper, a new modulation analysis and filtering method is proposed, in which an IF estimation approach based on conditional mean frequency (CMF) [7] is used for coherent modulation analysis and a carrier frequency tracking method is used to decompose multiple modulation components.

2. BACKGROUND

Start with a multi-component modulation signal model

$$x(t) = \sum_{n=1}^{N} s_n(t) = \sum_{n=1}^{N} m_n(t) \cdot c_n(t), \qquad (1)$$

where N is the number of components (e.g. harmonics); $m_n(t)$ and $c_n(t)$ are the modulator and carrier, respectively, for the nth component $s_n(t)$. The properties of the modulation signals are listed below

- $s_n(t)$ is a real or complex modulation component.
- $c_n(t)$ is a unimodular carrier signal, i.e. $c_n(t) = e^{j\phi_n(t)}$ if $c_n(t)$ is complex; $c_n(t) = \cos\phi_n(t)$ if $c_n(t)$ is real.
- $m_n(t)$ is a low-frequency modulator signal. It should be band-limited by the frequency range of $s_n(t)$.

2.1 Hilbert Envelope Detection

Hilbert envelope approaches such that the one used in modulation filtering studies by Drullman *et al* [8] and Atlas and Vinton [9] can be described as follows: For a real audio signal $x(t)$, the corresponding analytic signal $\hat{x}(t)$ can be obtained by Hilbert transform

$$\hat{x}(t) = H\{x(t)\} = |\hat{x}(t)| e^{j\phi(t)}, \quad (6)$$

where $\phi(t)$ is the phase of the analytic signal. $|\hat{x}(t)|$ is usually referred to as the *Hilbert envelope*. The modulator signal $m(t)$ and carrier signal $c(t)$ are then respectively given by

$$m(t) = |\hat{x}(t)|, \quad (7)$$

$$c(t) = \cos\phi(t). \quad (8)$$

The problem with Hilbert envelope approach is that forcing real and positive values for modulator signals induces phase discontinuities and thus causes the bandwidth of the modulator is much wider than the original modulation signal $x(t)$ [2, 5].

2.2 Coherent Demodulation

The coherent envelope detection was first introduced by Atlas and Janssen [5] for music source separation. In the coherent method, the modulator is no longer assumed to be real and positive. By allowing the modulator signal to be complex, phase discontinuities and the unlimited bandwidth issues now can be resolved. The work by Atlas and Janssen and subsequent studies [4, 6] have shown that modulation filtered audio signals with coherent approach offer significantly less distortion than Hilbert envelope based methods.

Without losing generality, we assume a target modulation signal $s(t)$ is an analytic signal. If the signal is real, the corresponding analytic signal can always be obtained through the Hilbert transform.

$$s(t) = A(t)e^{j\phi_m(t)} \cdot e^{j\phi_c(t)}, \quad (9)$$

where $A(t) = |s(t)|$. For a broadband signal with multiple components, a method to separate multiple components will be introduced in the next section. For now, we assume the signal has only one component. Supposing we have an approach to calculate the carrier signal phase term, the carrier signal $c(t)$ is given by

$$c(t) = e^{j\phi_c(t)}. \quad (10)$$

Then the modulator signal can be recovered by multiplying $s(t)$ by the inverse carrier phase term. This step is referred to as *coherent demodulation*.

$$m(t) = s(t) \cdot e^{-j\phi_c(t)} = A(t)e^{j\phi_m(t)}, \quad (11)$$

3. COHERENT MODULATION FILTERING

3.1 Instantaneous Frequency Estimate

It is a challenging problem to separate the modulation and carrier, because the separation of the phase term in (9) is not in general unique. In principle, for a *single-component* signal, an easy way to do this is through an instantaneous frequency (IF) estimate of the carrier signal. The instantaneous frequency $f_i(t)$ is defined as the derivate of the carrier signal phase

$$f_i(t) = \frac{d}{dt}\phi_c(t), \quad (12)$$

Once an IF estimate is obtained, the modulator can be calculated from (9) and modulation analysis or filtering can be performed. So IF estimation is a key to modulation analysis and filtering. In the proposed method, we will employ an IF estimation method introduced by Loughlin [7], which incorporates a natural band-limited condition. The IF estimation starts with time-frequency spectrographic analysis, which can be done through standard filter bank or short-time Fourier transform (STFT) analysis.

$$P(t,\omega) = \left|s(t) * h(t)e^{j\omega t}\right|^2 = \left|\int s(\tau)h(t-\tau)e^{-j\omega\tau}d\tau\right|^2, \quad (13)$$

where $s(t)$ is the signal; and $h(t)$ is the impulse response of a low-pass filter modulated by $e^{j\omega t}$ to each different frequency bin in the filter-bank interpretation of the spectrogram, or windowing function in STFT interpretation.

Please be noted that so far $s(t)$ is a single modulation component and insufficient for signals such as speech. Although filterbank analysis or SFTT is commonly used to separate different components in multi-component scenarios, here it is used only for the purpose of IF estimation. Multi-component separation will be discussed in the next section.

IF estimation is done via conditional mean frequency (CMF) of the spectrogram, which is defined as the first conditional moment of the distribution in frequency [10],

$$\langle\omega\rangle_t = \frac{\int \omega P(t,\omega)d\omega}{\int P(t,\omega)d\omega}. \quad (14)$$

Obviously, the CMF estimate depends upon the windowing function $h(t)$. Cohen and Lee [10] showed that, for a signal $x(t) = |x(t)|e^{j\phi(t)}$, the CMF of the spectrogram approaches the Hilbert IF $\phi'(t)$ when $h(t)$ becomes increasingly broadband; and it approaches the average frequency $\langle\omega\rangle$ when $h(t)$ becomes increasingly narrowband. When a proper $h(t)$ is chosen, the CMF yields a time-varying IF estimation that has meaningful physical interpretation [7, 11], i.e. the carrier frequency of a modulated single-component signal.

3.2 Multi-Component Decomposition

For broadband acoustic signals such as speech, there is usually more than one component. In previous methods, fixed filter-banks were used to decompose multiple components. The problem was that carrier signals could cross the subband boundaries. This is not a rare situation for acoustic signals. For example in speech signals, pitch harmonics act as the carrier signals. In normal speech, pitch changes will cause these boundaries to be crossed.

Furthermore, even though we can control subband width to make sure there is no more than one modulation component in the sub-band at any time, different modulation components may enter and leave a same specific sub-band at different time. As a result, the estimated modulator in that sub-band may belong to different components. If we do filtering over the estimated modulator, we actually accomplish modulation filtering across different components, which is not a well-defined behavior and produces unexpected results, e.g. acoustic distortions. Therefore, it is necessary to develop a method that is able to track the variations of the carrier signal as an alternative to fixed filter-banks.

In order to decompose multiple modulation components, we make following assumptions on carrier signals.

1. There is no frequency overlap between carrier components and between modulators of the neighboring carrier components.
2. Some *a priori* information of the signal is known, e.g. the number of carrier components and the maximum bandwidth of the modulators.
3. The stop-band response of the windowing function $h(t)$ in (13) is low enough to be ignored. Namely, the filterbank used in (13) has approximately ideal sidelobe behavior.

In order to avoid the previously-mentioned drawbacks of fixed filterbank method, a key concept is to track each of the carrier frequencies over time, and then accomplish time-varying band-pass filtering to separate each component. As shown in equation (14), the CMF is a statistical expectation of frequency of the modulation component at the time *t*, which indicates where the majority of the modulation component energy is located. Thus it is straightforward and theoretically justified to track the carriers by tracking the energy peaks on the spectrogram in (13).

Overall, the proposed modulation filtering approach consists of the following steps, also shown in figure 1.

1. Calculate a spectrogram with a properly chosen windowing function $h(t)$.
2. Determine the number of modulation components *n*.
3. Locate energy peaks on spectrogram and track them over time. The traces of the energy peaks give an initial IF estimates for each carrier component.
4. Calculate a finer IF estimate by using the CMF approach (eq. 14) within the bandwidth of the modulation component. In other words, instead of integrating over the full frequency range $(0,\pi]$, the integration is done only over the frequency range of the *n*th modulator $m_n(t)$.
5. Calculate the carrier phase signal by integrating the IF over time and calculate the modulator signals via (11)
6. Perform modulation filtering and reconstruct the time signal.

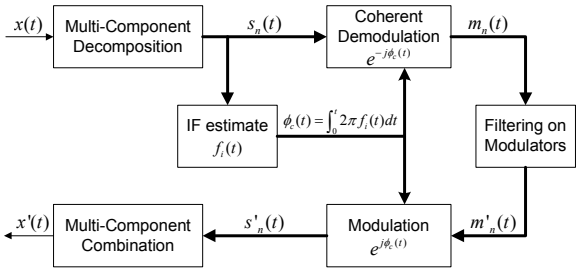

Figure 1. Proposed modulation filtering system

3.3 Frequency Shift Invariant

In our previous paper [6], three desired modulation filter properties are introduced. These properties are necessary for distortion-free processing. Of all the three properties, the superposition property and LTI property for modulator were achieved. However the frequency shift invariant property of modulation filtering was not achieved in any previous work.

Why is the frequency shift invariant property so important? This is because the whole purpose of modulation filtering is to do filtering on modulators while keep carriers unchanged.

To demonstrate the frequency shift invariant property for the new method, we start with a single component modulation signal. A time-frequency representation (e.g. STFT) has dual perspectives in time and frequency [12]. Specifically

$$\begin{aligned}\mathbf{X}(t,\omega) &= \int x(\tau)h(t-\tau)e^{-j\omega\tau}d\tau \\ &= \int X(\omega-\theta)H(\theta)e^{-j\theta t}d\theta\end{aligned}, \tag{15}$$

where $X(\omega)$ and $H(\omega)$ are Fourier transform of the signal $x(t)$ and $h(t)$, respectively. And the spectrogram in (14) is simply given by

$$P(t,\omega) = \left|\mathbf{X}(t,\omega)\right|^2 . \tag{16}$$

Obviously, for a frequency shifted signal $X'(\omega) = X(\omega+\omega_0)$, the STFT has frequency shift invariance.

$$\mathbf{X}'(t,\omega) = \mathbf{X}(t,\omega+\omega_0) . \tag{17}$$

The above result in equation (16) and (14)shows that the CMF satisfies the frequency shift invariant property.

$$\langle\omega'\rangle_t = \langle\omega\rangle_t + \omega_0 \tag{18}$$

For multi-component signals, given the condition of no frequency overlap between components, the frequency shift invariance property is achieved for each component, and thus achieved for the whole multi-component signal.

4. RESULTS

We will demonstrate the proposed method for two types of signals. One is a synthetic multi-component signal to verify how accurately the method can estimate carrier and modulator signals; and the other one is a speech signal which will show the distortion-free behavior during lowpass modulation filtering.

4.1 Multi-Component Synthetic Signals

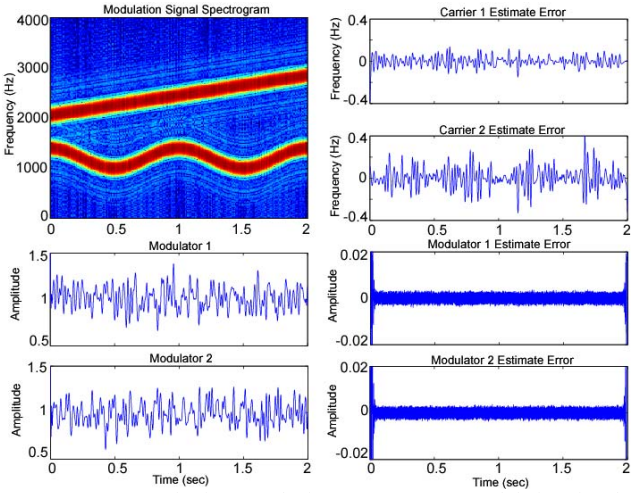

Figure 2. Test results for a synthetic two-component modulation signal.

As shown in figure 2, the synthetic signal consists of two components, one with a linear chirp carrier modulated by a low-pass filtered random signal and the other with a sine wave carrier modulated by a different low-pass filtered random signal. The bandwidth of the modulators is 100 Hz. A random frequency shift is added on the carrier signals to verify the frequency-shift - invariant property.

Both carrier and modulation estimate errors are plotted in figure 2. Errors are all very small. The carrier estimation errors are no greater than 0.4 Hz, a small fraction of the 4 kHz signal bandwidth. The modulator amplitude errors, other than edge effects, are less than 0.2 % of the average signal amplitude. When carrier signals are well defined and there is no overlap between the components, the proposed method can effectively separate multiple components, accurately estimate the carrier signals, and recover the modulator signals.

4.2 Speech Signals

In speech signals, each pitch harmonic is assumed to be a carrier which is modulated mostly by vocal tract variations. Although it may not be easy to track each harmonic frequency individually, we can take advantage of the harmonic structure use a pitch estimate to help track carrier frequencies.

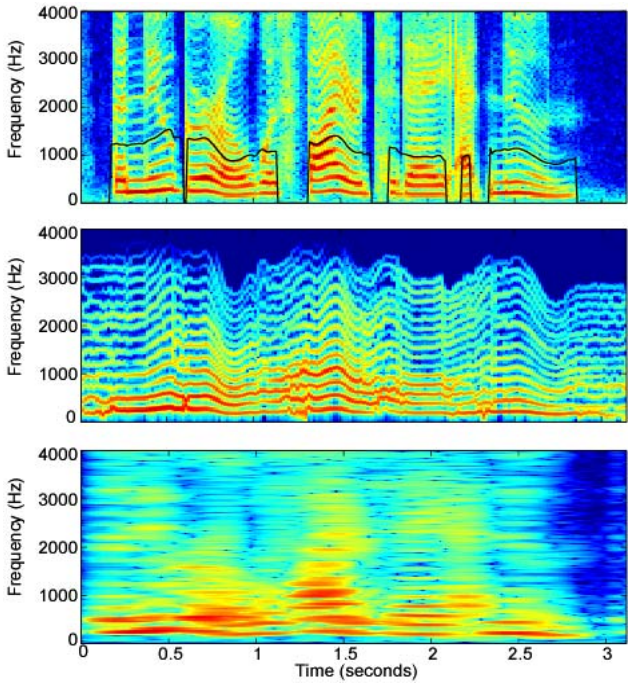

Figure 3. Low-pass modulation filtering results for a female voice "the new girl was fired today at noon." Upper panel shows spectrogram of the original signal, overlaid with the 5th pitch harmonic estimate (black line); middle panel shows the low-pass modulation filtered results with the propose method by the proposed method; lower panel shows the low-pass modulation filtered results with the fixed filter-bank method.

In the test, a female voice "the new girl was fired today at noon" is used. The ESPS normalized cross-correlation pitch estimator (http://www.speech.kth.se/snack) is used to estimate F_0. Based on the pitch range, the first 15 modulation components are modeled to approximately cover the frequency range of 0 ~ 4000 Hz. The pitch frequency indicates approximate positions of the carriers and the fine IF estimates are then obtained via the CMF based method (14). For silence or unvoiced speech, the pitch is interpolated for continuity of IF estimate.

For speech signals, we are more interested in how the modulation filtering modifies the speech signals. In the test, low-pass filtering was performed on the calculated modulators and compared with a fixed filter-bank based method. Note the same CMF based IF estimate method is used in both fixed filter-bank based method and the new proposed method.

As shown in the figure 3, after the modulation filtering with the new proposed method, the pitch harmonics in speech are perfectly maintained, while the vocal tract response is smoothed over time. On the other hand, with the fixed filter-bank based method, the vocal tract response is also smoothed, but the pitch harmonic structure is modified, particularly at the region where pitch changes fast (e.g. around 1.5 second). As a result, buzzing sound appears with extreme low-pass modulation filtering while the new proposed method produces distortion-free outputs.

5. DISCUSSION AND CONCLUSION

We have shown that the new proposed modulation filtering method is capable of estimating carriers and modulators accurately in modulation signals. The CMF based IF estimation approach along with a carrier frequency tracking algorithm offers a great tool for representing, analyzing, and processing acoustic signals. Particularly, owing to the pitch harmonic structure, the new proposed method works effectively with speech signals. A low-pass modulation filtering can achieve much temporal smoothness of the vocal tract response without change the pitch harmonic structure.

6. REFERENCES

[1] P. J. Loughlin and B. Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *J. Acoust. Soc. Am.*, vol. 100, pp. 1594-601, 1996.

[2] Q. Li and L. Atlas, "Over-Modulated AM-FM Decomposition," in *Proceedings of the SPIE*, Denver, 2004.

[3] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Am.*, vol. 110, pp. 1628-1640, 2001.

[4] S. M. Schimmel, K. Fitz, and L. E. Atlas, "Frequency Reassignment for Coherent Modulation Filtering," in *Proc. IEEE ICASSP*, Toulouse, France, 2006.

[5] L. Atlas and C. Janssen, "Coherent modulation spectral filtering for single-channel music source separation," in *Proc. IEEE ICASSP*, Philadelphia, 2005.

[6] Q. Li and L. Atlas, "Properties for Modulation Spectral Filtering," in *Proc. IEEE ICASSP*, Philadelphia, 2005.

[7] P. J. Loughlin, "Spectrographic measurement of instantaneous frequency and the time-dependent weighted average instantaneous frequency," *J. Acoust. Soc. Am.*, vol. 105, pp. 264-74, 1999.

[8] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, pp. 1053-1064, 1994.

[9] L. E. Atlas and M. S. Vinton, "Modulation frequency and efficient audio coding," in *Proceedings of the SPIE*, 2001.

[10] L. Cohen and C. Lee, "Instantaneous frequency, its standard deviation and multicomponent signals," *Proceedings of the SPIE*, vol. 975, pp. 186-208, 1989.

[11] P. J. Loughlin and B. Tacer, "Instantaneous frequency and the conditional mean frequency of a signal," *Signal Processing*, vol. 60, pp. 153-62, 1997.

[12] L. Cohen, "Time-Frequency Distributions - A Review," *Proceedings of the IEEE*, vol. 77, pp. 941-981, 1989.