COHERENT MODULATION SPECTRAL FILTERING FOR SINGLE-CHANNEL MUSIC SOURCE SEPARATION

*Les Atlas*and Christiaan Janssen***
*Department of Electrical Engineering, University of Washington
Seattle, WA 98195-2500, USA
**Fraunhofer IIS
Erlangen, Germany

ABSTRACT

Modulation spectral filtering, if effective and distortion-free, would offer a new tool for signal modification. Previous approaches to modulation spectral filtering, which made use of incoherent detection of real and positive modulating envelopes for each frequency sub-bands, have not offered effective and distortion-free signal modification. Based upon a recent observation that the modulating envelopes are potentially complex, coherent detection is instead proposed. Details are provided for accurate carrier estimation, and tests on both synthetic signals and music, show that modulation filtering is indeed distortion-free. The coherent modulation filtering method is applied to single-channel music sound source separation with promising results for music and other signal separation and modification applications.

1. INTRODUCTION

There is substantial evidence that many signals can be represented as low frequency modulators which modulate higher frequency carriers. Many researchers have observed that this concept, loosely called "modulation frequency," is useful for describing, representing, and modifying broadband acoustic signals. These observations have been the most common for, yet are not at all restricted to, speech and music signals. Modulation frequency representations usually consist of a transform of a one-dimensional broadband signal into a two dimensional joint frequency representation, where one dimension is typically standard acoustic frequency and the other dimension is a modulation frequency [1].

In this paper we focus on the concept of modulation filtering, which is the modification of a broadband signal's modulation frequency content. This filtering is intended to attenuate a signal's modulation content at a designed range of modulation frequencies, where these ranges can also be chosen as a function of acoustic frequency. Modulation filters potentially have a range of useful applications in signal enhancement and separation. For example, a well-designed modulation filter should be able to separate sounds which differ in their modulation content, such a percussive sound and a more tonal sound, even though the sounds overlap in time and regular acoustic frequency.

A number of modulation analysis and filtering techniques are described in literature, for example [2-6]. A problem with the existing methods, as reported by Ghitza [6] is that modulation filters show considerably less stop-band attenuation than what they are designed for. This lack of attenuation largely reflects the substantial distortion which comes about from an incorrect assumption of a real and positive modulation envelope [7] along with the corresponding incoherent methods of envelope detection. This distortion is severe enough to keep past approaches for modulation filtering from being suitable for high- or even low-fidelity applications, such as single-channel sound source separation.

In this paper, we propose as a foundation for high-fidelity modulation filtering, a coherent method of envelope detection for each acoustic frequency sub-band. In order to accurately coherently detect sub-band carriers and hence their envelopes, we developed a new instantaneous frequency estimator which has sufficiently low bias and variance. Our first tests, on synthetic signals, confirm that high-quality modulation filtering is possible. Our second tests, with results illustrated by spectrograms, on a single channel combination of a flute and castanets, confirm that correctly chosen fixed modulation filters provide substantial, though not complete, music source separation with no accompanying undesired artifact.

2. COHERENT APPROACH TO MODULATION SPECTRAL FILTERING

Previous approaches to modulation filtering, such as the Hilbert envelope approach of Drullman *et al* [2] or the magnitude envelope approach of Vinton *et al* [1], begin with filterbank or, equivalently, short-time transform analysis of the input signal, respectively. Each sub-band output is then detected to find an envelope. For example Drullman *et al* used a Hilbert envelope (magnitude of the analytic signal) and Vinton *et al* used a direct magnitude estimate of the envelope. Modulation filtering then consisted of some linear time-invariant filtering of each sub-band signal followed by a sum across sub-bands, to provide an output modulation filtered signal.

In order to better estimate the likely complex modulation envelope [7] and to reduce undesirable distortion during modulation filtering, we instead propose the coherent detection approach shown in figure 1. Referring to this figure, the

approach begins with sub-band decomposition similar to earlier approaches: A single-channel audio input signal $x[n]$ is separated into *K* acoustic frequency sub-bands via a filterbank or a short-time transform, such as a DFT. The change from previous approaches starts in the next step: Independently for each sub-band, typically after decimation, coherent detection consists of multiplication by the complex conjugate of the carrier estimate $e^{-j\omega_c k}$, resulting in a usually complex modulation envelope $m[k]$. This envelope is then filtered using standard linear time-invariant (LTI) techniques. Independent carrier estimates and LTI filters are also done for all *K* sub-bands, and these results are summed to produce the modulation filtered output $y[n]$.

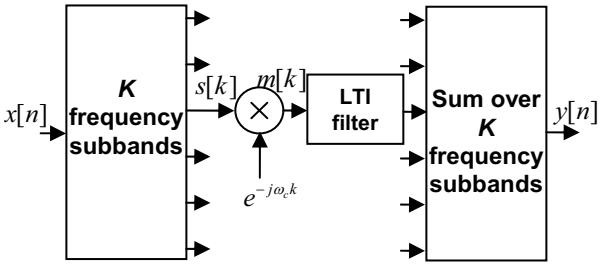

Figure 1. Coherent modulation detection and filtering. The coherent detection and modulation filtering operation is detailed for only one acoustic frequency sub-band.

As will be shown below, accurately estimating the carriers for each sub-band is the most difficult step of the above process.

3. CARRIER ESTIMATION

After ruling out more conventional carrier estimation approaches, such as phase-locked loops, as not having the time resolution needed, a differential detection approach was chosen. The proposed system is a substantial modification of the system proposed by Glas [8]. Glas' system was used for frequency-shift keying, where noise is a more serious issue than amplitude fluctuations. Our very different detection problem, which is for sub-band outputs for arbitrary audio inputs, has less inherent noise yet much more possible amplitude fluctuation than seen in radio systems. We modified this past approach to meet our need for an asymptotically unbiased and amplitude insensitive carrier estimator.

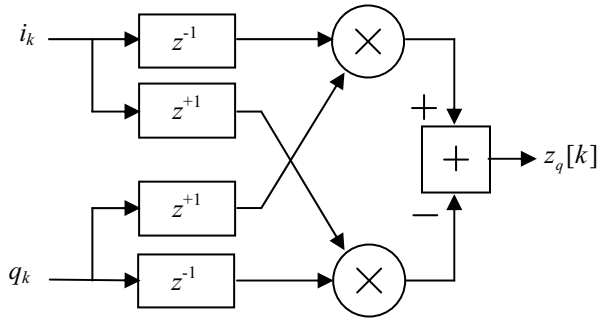

Figure 2. Quadrature part of the instantaneous frequency estimator. Using standard *z*-transform notation z^{-1} means a delay of 1 sample and z^{+1} means an advance of 1 sample.

This modified difference detector, as shown in figure 2 for only the quadrature output, first starts by decomposing the sub-band signal into its real (in-phase) and imaginary (quadrature) parts

$$i_k = \mathrm{Re}\, s[k],\ q_k = \mathrm{Im}\, s[k] \tag{1}$$

In order to reduce bias, we used a central difference formulation for the estimator. This estimator is

$$\begin{aligned} z_q[k] &= i_{k-1}q_{k+1} - q_{k-1}i_{k+1} \\ z_i[k] &= i_{k-1}i_{k+1} + q_{k-1}q_{k+1} \end{aligned} \tag{2}$$

where only the top part of eqn. 2 is shown in figure 2. These in-phase and quadrature parts can then be combined into an unnormalized phase estimate

$$z[k] = z_i[k] + jz_q[k] \tag{3}$$

To provide a unimodular phase-only estimate, the above quantity is normalized by its magnitude, resulting in the instantaneous phase estimate

$$\alpha[k] = \frac{z^*[k]}{|z[k]|} \tag{4}$$

This instantaneous phase estimates also uses a conjugate or, equivalently, the negative sign in the exponent, to demodulate the instantaneous frequency from the sub-band signal.

An instantaneous frequency estimate $W[k]$ can be recursively derived from the above phase estimate as

$$W[k] = W[k-1]\alpha[k] \tag{5}$$

where the initial conditions of the recursion are

$$W[-1] = 1,\ \ \mathrm{W}[0]{=}\alpha[0] \tag{6}$$

As seen in equation 6, this estimator becomes indeterminate when $|z[k]| = 0$, which occurs for vanishingly small signal levels. Thus we also add a condition of no change to the instantaneous frequency estimate, when the input is very small. Namely,

$$\alpha[k] = \alpha[k-1],\ \ \text{when}\ |z[k]| < \varepsilon \tag{7}$$

for some very small ε.

3.1. Bias of the Estimator

Consistent with our demodulation problem, we assume a narrowband sub-band signal model

$$s[k] = c[k]m[k] = e^{j\omega_c k}\left(m_c[k] + jm_s[k]\right) \tag{8}$$

where $c[k]$ is the assumed carrier and $m[k]$ is the assumed and desired modulator.

With this model, we can analyze the performance of the above equations 2-4. Applying the signal model from eqn. 8 to eqn. 2

$$\begin{aligned} z_q &= A\sin(2\omega_c) + B\cos(2\omega_c) \\ z_i &= A\cos(2\omega_c) - B\sin(2\omega_c) \end{aligned} \tag{9}$$

where ω_c is the carrier frequency for the sub-band and where *A* and *B* are the modulator terms

$$A = m_c[k-1]m_c[k+1] + m_s[k-1]m_s[k+1]$$
$$B = m_c[k-1]m_s[k+1] - m_s[k-1]m_c[k+1] \quad (10)$$

These modulator terms are related to the self correlation of the modulator signal. The *A* term gives a instantaneous estimation of the signal power, or self correlation, while the *B* term gives a measure of dissimilarity or cross-correlation between the phase and quadrature terms of this signal. In general, assuming low-pass symmetric modulation, the normalization process in eqn. 4 will tend to cancel term B. When this does not happen, this B term represents a bias in the angle estimation.

Increases in the size of ω_c will tend to give more importance to either the *A* or *B* terms in eqn. 9. This secondary source of bias is typically masked by the normalization process, since A tends to be some orders of magnitude higher than B.

There is the possibility to give concrete values for the bias for the worst-case condition of single-component overmodulation with a totally suppressed carrier. In this case our sub-band signal has only modulator terms, thus we can rewrite eqn. 8 as

$$s[k] = a \cdot e^{j\omega_1 k} + b \cdot e^{j\omega_2 k} \quad (11)$$

where

$$\omega_c = \frac{\omega_1 + \omega_2}{2} \text{ and } \omega_m = \frac{\omega_2 - \omega_1}{2} \quad (12)$$

Applying equations 2-3 gives

$$z[k] = e^{j2\omega_c} \cdot [A + j \cdot B] \quad (13)$$

where *A* and *B* now become

$$A = [(a^2 + b^2) \cdot \cos(2\omega_m)] + [2ab \cdot \cos(2\omega_m k)]$$
$$B = (a^2 - b^2)\sin(2\omega_m) \quad (14)$$

For the sake of clarity, the different terms of eqn. 14 have been grouped into a constant and a time-variant part. (*k* is the time index.) The instantaneous frequency estimator should give twice the carrier frequency as the angle of *z*[*k*]. In eqn. 13 can be observed that certain bias is introduced by the *A*+*jB* term. This bias evolves in time, but can be forced to be constant by smoothing the estimate, since the evolution has a cosinusoidal shape in time.

Calling the bias ξ, we can compute its expected value as

$$E\{\xi\} = \arctan\left(\frac{(a^2 - b^2)\sin(2\omega_m)}{(a^2 + b^2)\cos(2\omega_m)}\right) \quad (15)$$

This function depends on the parameters of the modulation. By inspection, it is clear that the bias is lower as the modulator is more symmetric about the carrier and the modulation frequency is lower. The bias effect also has a potentially positive effect. As observed empirically, when there are multiple potential carriers in a sub-band, as the asymmetry between the tones grows, the mean of the bias makes the estimator tend toward the higher power potential carrier. In other words, the estimator tends to follow the frequency of the highest power tone within the sub-band. This effect is increased with smoothing of the estimator.

4. RESULTS

In order to confirm the predicted performance of the proposed coherent modulation filtering system and carrier estimation approach, the overall system, with all sub-bands, was applied to both synthetic signals and natural music signals.

4.1. Synthetic Signals

As explained before, the carrier estimator becomes a tool for demodulating the signal, by demodulating it to the sub-band center, LTI filtering, and then restoring the carrier. Any filtering applied then to the sub-band signal will ideally affect only the modulator signals within each sub-band, leaving the carriers unchanged.

In order to illustrate these ideas, the next figures show the effect of applying an identical extreme low-pass LTI modulation filter within each sub-band.

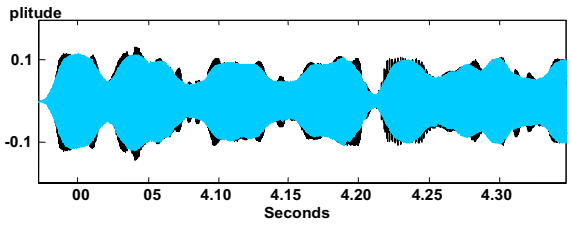

Figure 3. Time-domain plot of the effect of low-pass modulation filtering. Black is the original and light blue (grey for monochrome copies) is the modulation filtered signal.

Figure 3 shows the effect of low-pass filtering with a long frequency-sampled FIR filter designed using a Hamming window. While modulation filters were applied only within each sub-band, the reconstructed overall envelope of the signal is also clearly low-pass filtered, as intended.

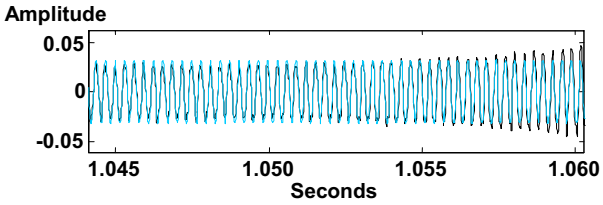

Figure 4. Zoomed-in time-domain plot of the effect of low-pass modulation filtering. Black is the original and light blue (grey for monochrome copies) is the modulation filtered signal.

Figure 4 shows a close-up of figure 3. In this case it can be seen that the phase and fine-time structure of the overall signal has been preserved. This indicates that the details of the carriers are correctly tracked and recovered for all sub-bands and that the final reconstruction was accurate.

4.2. Separation of Flute and Castanets

We applied the above modulation filtering technique to a single-channel sum of a high-fidelity flute and castanets recording. This trial confirmed that severe filtering in modulation, using our proposed coherent approach did not cause undesired artifact and that modulation filtering offers a promising new approach to the separation of signals with different dynamics.

Figure 5 shows coherently-calculated log magnitudes of DFT-based modulation spectra (with $K = 64$ subbands) for both a castanets signal and a flute signal, before they were summed to form the single-channel combination.

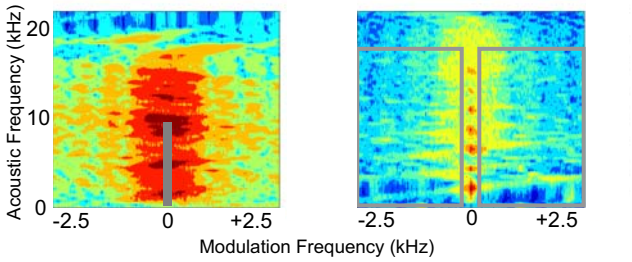

Figure 5. Modulation spectral plots of castanets (left) and a flute (right) with high-pass modulation filter and low-pass modulation filter stop-bands, respectively, overlaid with grey rectangles.

Figure 6 shows spectrograms of the unfiltered combination of the flute and castanets (top), the low-pass modulation filtered combination (middle), and the high-pass modulation filtered combination (bottom) using the modulation filters above. These filters emphasized the vibrato and sonorant quality of the flute or the percussive character of the castanets, respectively, resulting in 10-15 dB of signal separation without any undesired artifact.

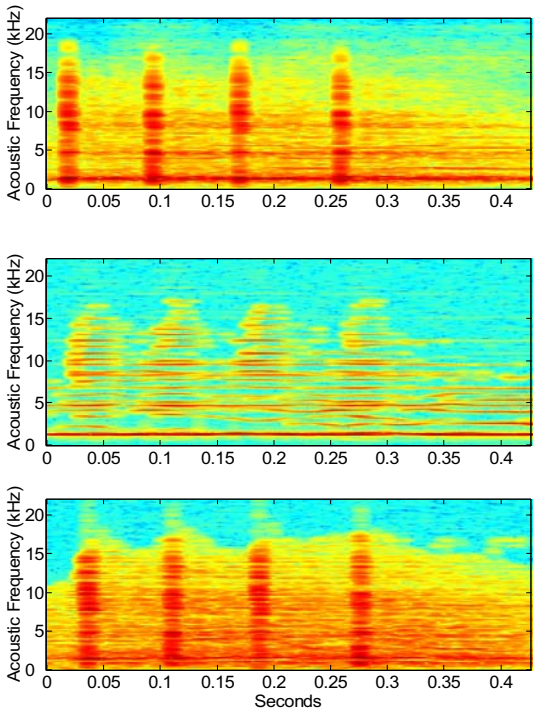

Figure 6. Spectrograms of the single-channel sum of the flute and castanets, before (top) and after processing.

5. CONCLUSIONS

Coherent modulation spectral filtering offers a new approach to the modification and separation of signals, based upon differences in dynamics. A coherent modulation spectral analysis and filtering approach was proposed, with single carrier detection, within each sub-band, central to the design. This carrier estimate was derived from a previous discriminant detector technique. Modifications were made to the previous technique to reduce and to control bias and to handle potentially small sub-band signal energies.

The performance of the proposed modulation filtering technique was confirmed on synthetic signals. Reconstruction across sub-bands showed essentially exact desired modification of the overall envelope, while signal fine structure was precisely maintained. Artifact-free and effective modulation low-pass and high-pass filtering was demonstrated on a sum of two high-quality music signals with differing dynamics. Low-pass in modulation strongly enhanced the sonorant source while high-pass in modulation strongly enhanced the percussive source.

Future work includes tests in other applications and comparisons to other promising techniques for carrier detection (e.g. [9]).

We acknowledge Sascha Disch and Juergen Herre of Fraunhofer IIS for their helpful discussions.

6. REFERENCES

[1] Mark S. Vinton, and Les E. Atlas, "A Scalable and Progressive Audio Codec*," ICASSP 2001*, pp. 3277–80.

[2] Rob Drullman, Joost M. Festen, and Reinier Plomp, "Effect of Temporal Envelope Smearing on Speech Reception," *Journal of the Acoustical Society of America,* Vol. 95, February 1994, pp. 1053–64.

[3] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of Speech with Filtered Time Trajectories of Spectral Envelopes", *Proc. ICSLP*, Vol. 4, pp. 2490–93, 1996.

[4] Steven Greenberg, and Brian E.D. Kingsbury, "The Modulation Spectrogram: in Pursuit of an Invariant Representation of Speech," *ICASSP 1997*, pp. 1647–50.

[5] A. Kusumoto, T. Arai, T. Kitamura, M. Takahasi, and Y. Murahara, "Modulation enhancement of speech as preprocessing for reverberant chambers with the hearing-impaired", *ICASSP 2000*, pp. 853–6.

[6] Oded Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception", *Journal of the Acoustical Society of America,* Vol. 110, September 2001, pp. 1628–40.

[7] Les Atlas, Qin Li, and Jeffrey Thompson, "Homomorphic Modulation Spectra", *ICASSP 2004*, pp. 761–4.

[8] Glas, J.P.F., "A differential FM detector for low-IF radios," Vehicular Technology Conference, 1999 (VTC 1999 - Fall. IEEE VTS 50th), Volume 2, 19-22 Sept. 1999, pp. 658-662.

[9] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive-instantaneous frequency of signals and application to speech," Journal of the Acoustical Society of America, vol. 105 (3), pp. 1912–1924, (March) 1999.