COHERENT ENVELOPE DETECTION FOR MODULATION FILTERING OF SPEECH

Steven Schimmel and Les Atlas
Department of Electrical Engineering, University of Washington, Seattle WA 98195, USA

ABSTRACT

Modulation filtering, which has been previously described as several related approaches to achieve modification of speech temporal dynamics, is shown to be less effective than intended. In particular, past Hilbert envelope approaches generate distortion which spreads across frequency sub-bands and modulation rejection is far from the amount intended. The source of this distortion is analyzed and a solution, based upon coherent envelope detection in each sub-band is proposed. This coherent approach is shown to be substantially more effective than conventional incoherent approaches on speech samples.

1. INTRODUCTION

There is substantial evidence that many natural 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 Fourier 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. For example, using a modulation filtering technique which will be described in the next section, Drullman *et al* [2] showed that the modulation frequency range of 4-16 Hz plays an important role in speech intelligibility. Modulation filters should also have a range of useful applications in signal enhancement and separation. For example, a well-designed modulation filter should presumably limit noise which occurs outside the range of modulation frequencies which are important for speech, thus increasing intelligibility. An understanding of modulation filtering is also crucial to related areas such as auditory psychophsics. Thus there is a need for theoretically sound methods for modulation analysis and filtering of speech signals.

A number of modulation analysis and filtering techniques are described in literature, for example [2-6]. A problem with many of the existing methods, as reported by Ghitza [5] is that modulation filters show considerably less stop-band attenuation than they are designed for. In our experiments, as will be shown in section 6, it was not uncommon to see only 2-13 dB effective stop-band attenuation for a modulation filter which was designed for 40 dB stop-band attenuation.

In this paper, we analyze the cause of this problem for the most common approach to modulation filtering, which is based on a decomposition of sub-bands outputs into a Hilbert envelope and Hilbert instantaneous phase. We then propose a new approach to modulation filtering, building upon the concept of a complex modulator proposed in [7]. We compare our approach to the previous Hilbert envelope approach and quantitatively show that it has substantially better stop-band attenuation. We conclude with an interpretation of these results and discuss problems which remain to be solved for carefully-defined modulation filtering.

2. HILBERT ENVELOPE APPROACH TO MODULATION FILTERING

Hilbert envelope approaches such as the one introduced by Drullman *et al* [2] can be described as follows. A waveform $x(t)$ is filtered through a filterbank characterized by a set of bandpass filters $h_k(t)$ to obtain the sub-band signals $x_k(t)$

$$x_k(t) = x(t) * h_k(t) \tag{1}$$

Using the Hilbert transform $H\{\cdot\}$, the analytic signal $x_{+,k}(t) = x_k(t) + jH\{x_k(t)\}$ of each sub-band is determined and each sub-band is decomposed into its envelope $a_k(t)$ and instantaneous phase (i.e. carrier) $c_k(t)$

$$a_k(t) = \left| x_{+,k}(t) \right| \tag{2}$$

$$c_k(t) = \cos \varphi_k(t) \tag{3}$$

such that

$$x_k(t) = a_k(t) c_k(t) \tag{4}$$

where $\varphi_k(t) = \arg[x_{+,k}(t)]$ is the phase signal as a continuous function of t. The envelope of each sub-band is subsequently filtered with a modulation filter $g(t)$

$$\hat{a}_k(t) = a_k(t) * g(t) \tag{5}$$

A broadband waveform is reconstructed by multiplying the modified envelope and the original carrier in each sub-band and summing across the sub-bands

$$\hat{x}_k(t) = \hat{a}_k(t) c_k(t) \tag{6}$$

$$\hat{x}(t) = \textstyle\sum_k \hat{x}_k(t) \tag{7}$$

It should be also noted that the carrier must be filtered by an all-pass filter $g_{ap}(t)$ with phase response that matches $g(t)$ for correct reconstruction.

There are several versions of the Hilbert approach described in the literature (e.g. [2,8-9]). They most notably differ in the use of either uniform or critical-band spaced sub-bands for the band-pass filters $h_k(t)$. The method by Drullman *et al* also differs from the others because it does not use the definition of the carrier from (3) directly. Instead, they define the modified sub-bands to be the original sub-bands multiplied by the ratio of the modified envelope and the original envelope

$$\hat{x}_k(t) = \frac{\hat{a}_k(t)}{a_k(t)} x_k(t) \tag{8}$$

However, it can be readily seen from (4) that (8) is equivalent to (6) except perhaps for differences in numerical stability.

3. LIMITATIONS OF THE HILBERT ENVELOPE

For the sake of minimum added distortion, it is important that the acoustic frequency bandwidth of the sub-bands after modification by modulation filtering is no greater than the bandwidth of the original sub-bands. The reason for this is that energy of the modified sub-band that falls outside the spectral region of the original sub-band acts as distortion in other sub-bands and is therefore undesired. The spilling of energy outside the sub-band also reduces the effective stop-band attenuation of a modulation filter, which is what we focus on in this paper. We will argue in this section that modulation filtering of the Hilbert envelope does not generally result in modified sub-bands that satisfy this bandwidth invariance property.

We are interested in the spectral content of the modified sub-band signal $\hat{x}_k(t)$. From (6) we have that the modified sub-band is the multiplication of the modified envelope and the original carrier signal in the time domain, which is the same as the convolution of their Fourier transforms in the frequency domain. In general, the bandwidth of the convolution of two signals in the frequency domain is the sum of the bandwidth of the two signals. Since the bandwidth of the modified envelope signal $\hat{a}_k(t)$ can be easily controlled by the choice of modulation filter $g(t)$, it is of less importance for our analysis of the band-width of the modified sub-band, and therefore we focus on the bandwidth of the carrier signal $c_k(t)$. Via a frequency domain representation of equation (4) we see that, by definition, the convolution of the Hilbert envelope and carrier in the frequency domain satisfies the bandwidth of the sub-band. However, it is well known that there is no physical reason that the Hilbert envelope itself is restricted to have the bandwidth of the signal it represents [10-11]. Because this Hilbert envelope is not band-limited, the only way that the convolution in frequency of the envelope and carrier is bandlimited is when the carrier is also not band-limited. In particular, it must contain a special wide-band structure of "cancellation terms" that exactly match and cancel the wide-band content of the envelope when they are convolved in a frequency domain representation of equation (4). After the Hilbert envelope is modulation filtered it no longer matches the special structure of the carrier. As a consequence, the modified sub-band will typically have greater band-width than the original sub-band.

4. THE NEED FOR COHERENT CARRIER DETECTION

We conclude that in order to achieve small bandwidth in the reconstructed sub-band, the detected carrier must be a narrow-band or ideally, a monochromatic signal. Moreover, the carrier must depend on the input signal in a coherent way. For example, if the carrier is an incoherent pure tone at the frequency of the center of the sub-band, the envelope detection operation is nothing more than demodulating the sub-band by that frequency. Since that is a linear operation, the complete modulation filtering system, considered over all sub-bands, would be a periodic extension in frequency of a linear time-invariant (LTI) system, which still is LTI and therefore does not qualify as the novel form of modulation analysis and filtering which is intended.

Hence, we must use a coherently detected carrier signal in the carrier-envelope decomposition of each sub-band. We use the term coherent here to indicate a carrier signal that has an instantaneous phase that is in some way related to the phase signal $\varphi_k(t)$ of the sub-band. The optimal carrier, in the sense that it is the most narrow-band coherently detected carrier, is the monochromatic carrier at the "average" frequency of the sub-band. We refer to this frequency ω_r as the midband frequency, a concept and term first used by Rice ([12] p. 75) to describe a specific frequency in a sub-band that is not necessarily at the center of the sub-band. In this work, we define the average frequency of a sub-band to be the frequency ω_r (with initial phase φ_r) such that the phase signal

$$\theta_k(t) = \varphi_k(t) - \omega_r t - \varphi_r \tag{9}$$

has zero mean. It is easy to verify that there exists a unique pair (ω_r, φ_r) that satisfies this condition.

Since signals such as speech are far from stationary or cyclostationary for long durations in time, it is not reasonable to use a single midband frequency estimate for the entire duration of a (possibly long) input speech signal. On the other hand, from arguments in the previous section, we cannot choose a carrier that has exactly the same phase signal as the sub-band, because that would be the previous Hilbert carrier which has unrestricted bandwidth. Instead we suggest using a carrier signal that approaches the true phase signal to some degree, but also is narrowband. It can be viewed as an estimate of Rice's midband frequency that slowly varies over time. The rate of change of the estimate can be varied by a smoothness parameter, and the "best" rate will depend on the lack of stationarity of the input signal.

As we will see from the definition of our new coherent approach in the next section, both the envelope that corresponds to the detected carrier and the carrier will in general be complex. Although this may appear unfamiliar at first, the use of complex envelopes and their necessity for modulation frequency analysis and filtering was justified in recent work [7].

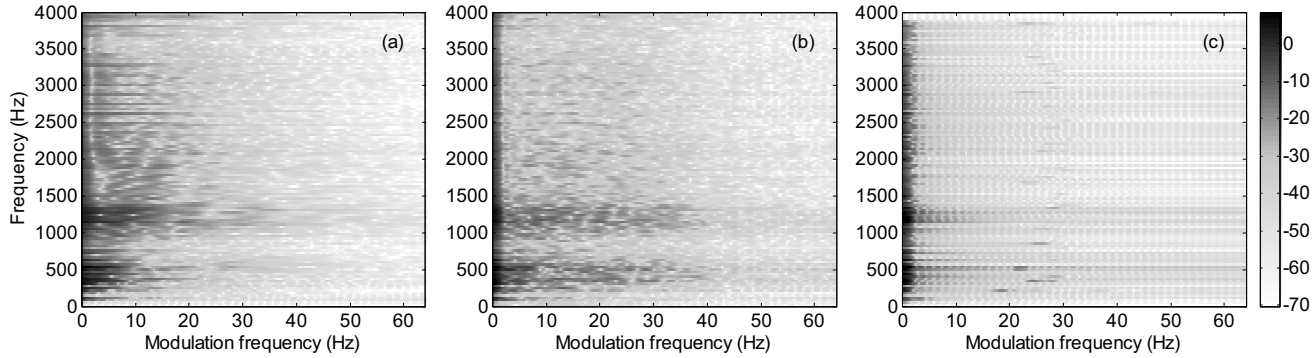

Figure 1 (a) Modulation spectrogram of the original speech signal; (b) Modulation spectrogram of the speech signal filtered with the conventional Hilbert approach; (c) Modulation spectrogram of the speech signal filtered with the new coherent approach.

5. PROPOSED COHERENT APPROACH

Our approach is simplest to define by combining (3) and (6) to find that we can write the analytic signal of a sub-band as

$$x_{+,k}(t) = a_k(t)\exp[j\varphi_k(t)] \tag{10}$$

This implies that the Hilbert envelope is not only defined as the magnitude of the analytic signal, but that it also can be found via coherent detection to remove the effect of an assumed carrier

$$a_k(t) = x_{+,k}(t)\exp[-j\varphi_k(t)] \tag{11}$$

For our coherent detector, we substitute the zero mean phase signal $\theta_k(t)$ as defined in (9) into (11) to get

$$a_k(t) = x_{+,k}(t)\exp[-j(\theta_k(t) + \omega_r t + \varphi_r)] \tag{12}$$

By smoothing (i.e. band limiting) the phase signal $\theta_k(t)$, the slow rate of change of the midband frequency estimate can be controlled. With no smoothing of $\theta_k(t)$, the detected envelope equals the Hilbert envelope, and with full smoothing of $\theta_k(t)$, the detector returns the complex envelope $a_k(t)$ corresponding to the monochromatic carrier at the midband frequency. Let $\tilde{\varphi}_k(t) = \theta_k(t) * h_{lp}(t) + \omega_r t + \varphi_0$ therefore be a smoothed version of the phase signal that has been smoothed by the low-pass filter $h_{lp}(t)$. The decomposition of a sub-band into an envelope and carrier in our coherent complex envelope approach is then

$$a_k(t) = x_{+,k}(t)\exp[-j\tilde{\varphi}_k(t)] \tag{13}$$

$$c_k(t) = \exp[j\tilde{\varphi}_k(t)] \tag{14}$$

These equations replace equations (2) and (3) to form the complete modulation filtering approach. In our approach, the low-pass filter $h_{lp}(t)$, and in particular its stop-band cut-off frequency *lp*, gives control over the distortion-free performance and effective stop-band attenuation of modulation filters by limiting the band-width of the carrier signal.

6. RESULTS

All of the results presented in this section are based on modulation filtering of speech, either with the standard Hilbert approach or with our proposed coherent approach. In both approaches we used the same perfect reconstruction filterbank, which consisted of sub-band filters with a sine-squared frequency response and a bandwidth of 64 Hz. There was 50% overlap in frequency between the sub-bands. This narrow and uniform bandwidth was chosen so that resolution in modulation frequency would be high. Similar conclusions should be possible for non-uniform and/or broader bandwidth sub-bands. In order to best illustrate non-ideal effects, we applied the same severe low-pass modulation filter to the envelopes of all sub-bands. The modulation filter had a cut-off frequency of 2 Hz with 40 dB stop-band attenuation.

Figure 1 shows the modulation spectrogram of a short speech signal before modulation filtering (1a), as well as after modulation filtering using the Hilbert approach (1b) and after filtering using our proposed approach (1c). We used the approached described in [1], basically a magnitude spectrum of each frequency index in time of a standard magnitude spectrogram, to calculate the magnitudes of the modulation spectrograms seen in figure 1. This analysis was also used for the experiments reported below. The figure shows that the new approach suppresses modulation frequencies substantially better than the standard Hilbert approach, but some regular distortion is still visible. For this figure, the phase signal in the new approach was completely smoothed. Smoothing the phase signal less severely will reduce this distortion, but, as will be seen below, with the cost of less rejection in modulation filter stopbands.

To quantify the effective stop-band attenuation of the new approach for different levels of smoothing of the phase signal compared to the Hilbert approach, we tested the effectiveness of both approaches on 10 short speech signals sampled at 8 kHz. The samples, ranging in duration from 625 to 950 milliseconds, contain speech from a male speaker saying an isolated letter or digit. All processed speech samples showed a lack of ideal modulation low-pass behavior to varying degree. In order to summarize this in one figure, we averaged the effective modulation frequency responses of each approach across all acoustic frequencies sub-bands and across all speech samples. For modulation filtering, all speech signals were first separated into sub-bands using the filterbank described above. The sub-band envelopes were detected using either the Hilbert approach, the coherent approach with complete phase smoothing over the signal duration, or the coherent approach with a 4 Hz low-pass filter applied to the phase signal. The phase smoothing filter was designed with 40 dB attenuation in the stopband. Next, the detected envelopes were low-pass filtered using the severe low-pass modulation filter described above, and recombined with their carrier signals.

Finally, the modulation filtered single-channel speech signals were reconstructed by summing the filtered sub-bands.

For each approach, the effective modulation frequency response was measured as the ratio of the energy in the modulation spectrogram of the original signal to the energy in the modulation spectrogram of the modified signal, averaged over all acoustic frequencies and all modified speech signals. The combined result, shown in Figure 2, is the effective modulation frequency response of the severe low-pass modulation filter for each approach. This figure shows that the new coherent approach achieves substantially better stop-band attenuation for both choices for the smoothing filter $h_{lp}(t)$. The reduced suppression for modulation frequencies around 32 Hz is caused by the 32 Hz spacing of the sub-bands of the filterbank. This artifact is audible when the phase signal is completely smoothed, but less audible for the other choice of the low-pass phase smoothing filter.

7. CONCLUSIONS

Modulation filtering, if accurate and distortion-free in its effect, would be a useful new tool for speech and signal processing. Previous work in this area has been shown to not achieve these goals. We have confirmed and identified the details of why previous Hilbert envelope approaches cause undesirable distortion and strongly detract from the desired modulation filtering effect. From this understanding and previous results showing the need for a complex modulation envelope, we have justified the need for coherent carrier estimation. This paper represents a first attempt at coherent carrier estimation within the context of modulation decomposition and filtering. We also are the first to propose that modulation frequency filtering effectiveness be measured in terms of amount of modulation frequency rejection.

Measured results on speech, which compared our coherent approach to the more conventional Hilbert approach, confirmed that the conventional approach has only a weak stop-band rejection of about 3-14 dB and that the coherent approach can reject 6-28 dB. Since the proposed coherent approach was for an ideal rejection of 40 dB, problems still remain. In particular, accurate carrier phase estimates, appropriate amount of smoothing of carrier phase estimates, and possible distortion during reconstruction across sub-bands are not yet fully understood. However, future careful study of the proposed coherent and other carrier estimation techniques (e.g. [13]) is expected to result in ideally effective and distortion-free modulation filtering.

This research was supported by the Washington Research Foundation. We acknowledge Prof. Bishnu Atal and Qin Li of the University of Washington for their helpful discussions.

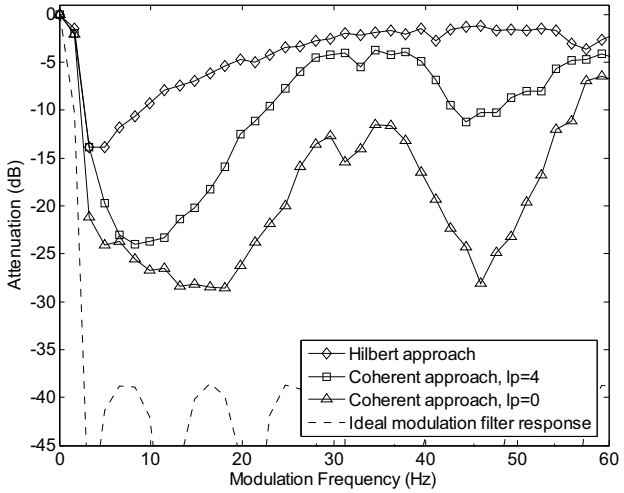

Figure 2 Effective modulation frequency responses for the Hilbert approach and the coherent approach with two levels of smoothing. The parameter *lp* specifies the cut-off frequency of the low-pass filter used.

8. 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] 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.

[6] Jeffrey Thompson, and Les Atlas, "A Non-Uniform Modulation Transform for Audio Coding with Increased Time Resolution," *ICASSP 2003*, pp. 397–400.

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

[8] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception", *Nature*, Vol. 416, March 2002, pp. 87–90.

[9] 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.

[10] J. Dugundji, "Envelopes and Pre-Envelopes of Real Waveforms", *IRE Trans. on Information Theory*, Vol. 4, pp. 53–7, 1958.

[11] Bernard Picinbono, "On Instanteneous Amplitude and Phase of Signals", *IEEE Trans. on Signal Processing*, Vol. 45, No. 3, March 1997, pp. 552–60.

[12] S.O. Rice, "Mathematical Analysis of Random Noise," *Bell Sys. Tech. J.*, Vol. 24, No. 1, Jan. 1945, pp. 46–156.

[13] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Transactions on Speech and Audio Processing*, **8**, pp. 240-54, 2000.