A MODULATION VIEW OF AUDIO PROCESSING FOR REDUCING AUDIBLE ARTIFACTS

David V. Anderson
Georgia Institute of Technology, Atlanta, Georgia, USA. E-mail: anderson@gatech.edu

ABSTRACT
When manipulating audio signals, frequency decomposition, gain functions, and temporal behavior can be understood in terms of modulation effects and their perceptual impact. For example, audio enhancement or dynamic range control both rely on spectral gain functions that fluctuate rapidly enough to adapt to changing signal characteristics. However, rapid fluctuations in gain effectively modulate the signal, resulting in perceptual artifacts. This paper discusses the perceptual impact of modulating gain and trade-off between temporal responsiveness and perceptibility. We demonstrate that, in many situations, it is best to apply gains in subbands that are similar to critical bands in the human auditory system and that the gain function in each subband should be low-pass filtered with a cutoff frequency proportional to the subband bandwidth.

Index Terms— Speech enhancement, modulation, dynamic-range compression, FFT, cochlear filtering

1. INTRODUCTION
Anyone who has developed a noise suppression filter using spectral subtraction, adaptive Wiener filtering, or some related approach knows that one of the primary difficulties is reducing the distortion and artifacts introduced by the filter itself. These filters are applied in the frequency domain by estimating the optimal filter gain, applying it to the frequency-domain noisy signal, and constructing the “clean” signal in the time domain using an overlap-add technique.

A common artifact resulting from such filtering is called musical noise—so named because it often sounds like brief random notes or bird chirps. The artifacts are often blamed on imperfect estimates of the optimal filter and so the filter estimates are often averaged over time to provide better optimal filter estimates. However, these artifacts are introduced by rapid changes in the filter and not just an inaccurate estimate of the optimal filter: a fixed filter, even one with an undesirable frequency response, does not introduce the distortions usually associated with noise reduction such as musical noise. These distortions are only introduced as the filter changes rapidly between frames (see also [1]). Smoothing the filter between frames drastically reduces musical noise and related effects but may result in other artifacts such as ghost speech or reduced noise-suppression. To combat these problems, many noise-suppression systems utilize adaptive temporal averaging (e.g., see [2]).

Systems that perform dynamic range compression (DRC) also use time–varying gain. Examples include compression hearing aids, systems that use adaptive gains to reduce the peak–to–RMS ratio in a signal, and systems that equalize loudness levels over time and frequency. With such systems it is particularly easy to investigate the effects of time–varying gain because the situation is not complicated by the presence of additive noise. This paper uses time–varying gain for fast DRC (at sub-phoneme time scales) as a vehicle to discuss time–varying gain and associated artifacts. Although this paper does not present results for noise suppression, the concepts apply to noise suppression and any other audio processing system that employs subband time-varying gain.

2. TIME-VARIANT SYSTEMS
The goal of this paper is to develop a system whereby time-varying gains may be applied to a signal to achieve desired effects without using adaptive temporal averaging and without producing any other detectable changes. For example, performing fast dynamic-range compression or signal enhancement without any other perceivable artifact.

When an applied gain changes slowly enough, a listener may notice that the “volume” is changing but not associate it with any particular artifacts. The problems occur with rapidly varying gain—in this case, the gain acts as an amplitude modulation function. Amplitude modulation generally causes a spreading of the frequencies in a signal which can result in significant cues that the auditory system can detect such as a roughness or other artifacts [3]. The amount of spreading is a bounded function of the modulator bandwidth and can be calculated. The maximum frequency deviation in the modulated signal is equal to the highest frequency in the modulating gain signal. The modulation product terms are likely the significant cues that the auditory system uses to identify the variation in gain [4].

2.1. Critical Bands and Subband Representations
The cochlear (inner ear) is often regarded as performing a frequency transform on incoming audio signals. This is true to some degree, but the transform is not linear and not time-invariant. However, if we proceed with caution, some aspects of earlier auditory filtering can be approximated as linear operations. Critical bands are a concept that captures some of the gross aspects of cochlear filtering. Generally speaking, signals within a critical band may mask each other while signals in separate critical bands have a minimal masking effect on each other. Furthermore, a given amount of energy distributed within a single critical band yields the same loudness percept regardless of how it is distributed [5]. If that same energy is spread into a second critical bands, the perceived loudness increases.

These two concepts play an important role in the following analysis.

Before proceeding further, it will useful to define a simple signal representation that lends itself to perceptual analysis. For this representation the signal is divided into subbands indexed by $k$; and then each subband is further decomposed into a product of an envelope (which carries the loudness information) and a “carrier” signal of nearly constant power [6]. This signal representation can be applied to auditory analysis by making the signal subbands roughly equal in bandwidth to the critical bands in the ear. With this representation, the loudness of the signal perceived in any particular critical band of the ear may be primarily controlled by operating only on the envelope in that band. In particular, the acoustic signal, $s(t)$, is written as

$$s(t) = \sum_k s_k(t) = \sum_k e_k(t) u_k(t)$$

(1)
where \( v_k(t) \) is a carrier signal with nearly constant power; and \( e_k(t) \) represents the envelope variation over time (see Figure 1). Figure 2 shows \( s_k(t) = e_k(t)v_k(t) \) for a single band from a sentence spoken by a female.

![Diagram of Filter Bank](image)

**Fig. 1.** The signal is divided into subbands and each subband signal is represented as a product of a slowly varying envelope, \( e_k(t) \), and a nearly constant-energy vibration signal, \( v_k(t) \). Each subband is then compressed via a time–varying gain and recombined.

![Graph of Filtered Speech Signal](image)

**Fig. 2.** The gray area is \( s_k(t) = e_k(t)v_k(t) \) for a single band taken from a speech signal. The dark line is the envelope, \( e_k(t) \).

### 2.2. Minimizing Perceived Distortion

To minimize the modulation effects of time-varying gain, constraints are placed on the modulating gain function to either decrease the average modulation depth or to decrease the highest modulation frequency. An example of a technique that decreases the average modulation depth is presented in [7] where the attenuation in limited to the minimum required to ensure that the noise is masked. Decreasing the modulation frequency is a very common approach and is usually implemented by smoothing the gain or otherwise limiting its rate of change [8, 2].

#### 2.2.1. Perception of Modulated Signals

If a signal changes pitch or if new frequencies are audible after the application of a gain function, an audible artifact has been introduced. Minimizing the maximum modulation frequency can reduce the amount of pitch shift and/or the generation of new pitches. Note, that for slowly modulated waveforms, e.g. around 5 Hz, the effect is perceived as a fluctuating loudness. As the modulating frequency increases, it the effect may be heard as a roughness [5]. In general, when the gain is responsive to speech signal characteristics, the modulation is fast enough to produce the “roughness” artifact rather than the fluctuating loudness.

One approach to selecting the maximum modulating frequency to minimize artifacts is to take inspiration from auditory frequency JNDs (just noticeable differences). Frequency JNDs are approximately logarithmic in frequency indicating that humans can resolve low frequencies better than high frequencies. Thus, it could be possible to limit any gain modulation to be lower in bandwidth than the frequency JND corresponding to the lowest subband to which it is applied, ensuring that modulation products are indistinguishable in frequency from the original waveform. However, frequency JNDs are based on measurements with pure tones and the applicability to speech signals is not established. Accordingly, we have found that the frequency JNDs tend to limit the modulating gain much more than necessary for complex stimuli.

As discussed in Section 2.1, if the modulation spreads energy within a single critical band, the overall loudness of the signal in that band remains essentially the same. However, if the modulation spreads energy into adjacent critical bands, the perceived loudness changes. Off-frequency listening refers to the practice of listening for changes in a signal by detecting changes in nearby frequencies; this is usually sub-conscious. A simple example of this occurs when there is a small discontinuity in a sinusoidal stimulus; the resulting click, although it may contain very little energy, may be quite audible.

Off-frequency listening provides another metric by which to constrain a modulating gain. That is, simply constrain the gain so that most of the spectral artifacts remain within the same critical band as the modulated signal. This is accomplished by letting the \( s_k(t) \) signals be the outputs of critical bandwidth filters. Under that assumption, the lower frequency bands have \( v_k(t) \) waveforms that are almost sinusoidal while higher bands may have bandwidths of over 1000 Hz. The subband envelopes \( e_k(t) \) are lowpass in nature and have bandwidths proportional to, but less than, the bandwidths of the associated critical band signals. The gain may then be chosen as a function of the envelope, \( e_k(t) \). The modulation terms are controlled by either limiting the bandwidth of \( e_k(t) \) (if the gain function is approximately linear) or the output gain itself. The bandwidth of \( e_k(t) \) can be limited naturally when estimating the envelope, eliminating the need for further filtering. For typical audio and speech operations we have found that limiting to the bandwidth of \( e_k(t) \) to between \( bw_k/8 \) and \( bw_k/2 \) is sufficient where \( bw_k \) is the bandwidth of the \( k \)th subband.

#### 2.2.2. Problems with FFT-Based Systems

Any method that changes the gain faster than suggested in the previous paragraph would cause significant energy to appear in adjacent auditory filters which is perceived as a distortion product. Any method that changes the gain more slowly may be too slow to adjust for rapid signal changes; for example, if a sudden loud noise occurs at the input to a DRC, the gain must adapt quickly enough so that excessive gain is not applied to the loud signal.

Using any fixed-bandwidth analysis filter presents a problem because the bandwidths of the envelopes for all bands of a fixed-bandwidth system are the same. Thus, the maximum allowable modulation frequency is determined by the lowest-frequency band; and the gain modulation for higher frequency bands would not be as rapid as it could be with a critical-band analysis. Another way to view this is that the length of the basis function of the FFT limits the rapidity of response across all frequencies. By using an analysis method that is matched to the human auditory system, i.e. critical band filters, it is possible to maximize the responsiveness of the time-varying gain across all frequencies subject to the modulation limits.
2.2.3. Modulation and Speech

When processing speech signals, such as for noise suppression, the modulation signal decomposition in equation 1 is particularly useful. In this case, $e_k(t)$ can be additionally constrained by a knowledge of the characteristics of speech. Namely, at higher frequencies, $e_k(t)$ may be low-pass filtered to allow the speech signal to pass (it’s highest modulation frequency is limited by the speech articulators) while removing higher-frequency modulation due to background noise.

3. EXPERIMENTAL DEMONSTRATION

The effect of time-varying gain is demonstrated on several signals designed to demonstrate the transient and steady-state behavior. The particular gain function chosen for this demonstration is one designed to reduce dynamic range, such as might be used in hearing compensation or for adding noise immunity to an audio signal. The goal is to adapt the gain slowly enough not to cause audible distortion and fast enough so that the dynamic range of the signal is kept nearly within some range even when there are sharp discontinuities in signal level. For this scenario, it is important to operate on subband signals since it is desirable to do DRC across the spectrum. The gain function used to accomplish the DRC is

$$g_k(e_k(n)) = \min \left( \beta_k e_k^{\alpha_k-1}(n), g_{max} \right) \quad (2)$$

where $\alpha_k$ and $\beta_k$ are parameters that describe the severity and normalization of the dynamic range compression. A limit, $g_{max}$, is placed on the gain so that silent or nearly silent portions of the signal are not given excessive gain.

3.1. Linear Frequency Decomposition

Frequency-domain processing is often implemented in the context of a linear frequency decomposition using transforms such as the DFT (FFT) or DCT. Linear-frequency subband DRC is demonstrated here using a FFT–based filter-bank in which decimation is performed in the subbands by hopping the DFT frame in such a way that the subbands are not critically sampled, thereby avoiding aliasing issues. The envelope $e_k(n)$ is calculated in two different ways,

1. $e_k(n)$ is the magnitude of the complex value of bin $k$.
2. $e_k(n)$ is the magnitude of the complex value of bin $k$ smoothed over multiple frames to effectively slow down the gain adaptation.

Several DFT window lengths (5, 10, and 20 msec) were used with method (1) above for finding $e_k(n)$. For method (2) above, the analysis DFT window length was 2.5 msec with an effective 20 msec time–constant for the $e_k(n)$ smoothing. The overlap of analysis frames was 87.5%—the DFT frame was hopped by the window length divided by eight and the outputs were interpolated using a windowed overlap-add procedure with raised cosine windows. The DFT length used is twice the window length to minimize circular convolution problems when applying a frequency domain gain. One problem that can arise with DRC using DFT analysis is that the sidebands may be given excessive gain. This problem is addressed by using a Hamming window to reduce the sidebands and by limiting the maximum gain as shown in eq. 2.

3.2. Critical-Band Frequency Decomposition

The critical-band frequency decomposition is performed using FIR filter-banks without decimation (thus avoiding any possible questions of aliasing for this analysis). The envelope of each subband is then estimated as

$$e_k(n) = \sum_{m=-[T_k/2]}^{[T_k/2]} u_k(m)s_k^2(n - m). \quad (3)$$

The time-constant $T_k$ is chosen as $T_k \approx 8f_s/bw_k$ where $bw_k$ is the bandwidth of the $k$th filter. Note, another option would be to have a fixed time–constant for each band but that defeats the advantage that the constant–Q method presents of being able to adapt quickly to transients while not distorting the frequency content in the lower–frequency bands. Also, notice that the envelope calculation is non-causal; this may not be feasible in practice but is done here for several reasons. First, it is possible to get fairly good envelope detectors with minimal delay but it is beyond the scope of this article to investigate the appropriateness of various methods. Second, the block processing of the FFTs used in the linear frequency decomposition provides an excellent estimate of the energy at each frequency; to ensure a fair comparison, a similarly good estimate is important for the critical–band case.

3.3. DRC Demonstration

Several pure–tone signals with gain discontinuities and complex signals were processed in order to isolate and demonstrate the artifacts generated using the various processing techniques described above. These signals are:

- A single 100 Hz tone with an amplitude discontinuity.

$$x_1(k) = \begin{cases} 0.2 \cos(2\pi100k/f_s) & \text{if } k \leq 8000, \\ 0.8 \cos(2\pi100k/f_s) & \text{if } k > 8000, \end{cases} \quad (4)$$

- A single 2300 Hz tone, $x_2(k)$, similarly defined.

- A complex tone, $x_3(k)$, similar to the above but containing the following frequencies: 100, 110, 525, 1300, 2700.

Signals were processed to do fast DRC on subbands using three filtering methods: narrowband compression (using a DFT), narrowband compression with 20 millisecond temporal envelope smoothing, and critical bands approximated using constant-Q filters. For all methods, a maximum gain limit, $g_{max}$, was set at 30 dB to reduce excessive gain for residual signals (or, in the case of speech, to prevent between–word background noise from receiving excessive amplification). All of the signals were sampled at 8 kHz. The narrowband compression used 32 complex subbands while the compression based on cochlear filtering used 15 subbands each with a Q of about 3 (1/3–octave filters) to approximate the critical bands.

Spectrograms comparing the results are shown in Fig. 3 (note that all spectrograms are scaled the same so that comparisons between them are meaningful). There are several things to notice here. The two narrowband methods produce audible modulations terms in every case. The constant-Q approach causes some spreading of the transient at lower frequencies (most noticeable in Fig. 3a); however, this spreading corresponds to 1-2 periods of the signal at those frequencies so it is most likely not perceptually relevant.
3.3.1. DRC Listening Results

Fifteen listeners listened to 12 sentences, each processed using three different subband DRC approaches as described above. The sound was presented through headphones and subjects were allowed to play sounds as many times as desired before ranking the processed sentences according to preference. The presentations were randomized but all subjects listened to all conditions. The sentences were randomly chosen, with equal male and female talkers, from the TIMIT database. For each sentence, the audio files processed by the three methods were ranked in order of preference with 3=best and 1=worst.

The critical-band processing was nearly universally preferred, achieving an average score of 2.99. The smoothed-gain FFT processing had an average score of 2.02, and the unsmoothed-gain FFT processing was almost universally regarded as the poorest quality with a score of 1.02.

4. CONCLUSIONS

This paper shows that applying frequency–dependent, time–varying gain to audio may be better performed using critical-band filters rather than using constant–bandwidth subbands (e.g. FFT). In particular, we argue that perceptual distortion can be understood in terms of modulation effects and that the distortion is significantly less when minimizing modulation artifacts.

5. REFERENCES