Noise Suppression Based on Extending a Speech-Dominated Modulation Band

Tiago H. Falk[1], *Svante Stadler*[2], *W. Bastiaan Kleijn*[2], *and Wai-Yip Chan*[1]

[1]Department of Electrical and Computer Engineering, Queen's University, Canada
[2]School of Electrical Engineering, KTH - Royal Institute of Technology, Sweden
Email: {falkt, chan}@ee.queensu.ca, {svante.stadler, bastiaan.kleijn}@ee.kth.se

10.21437/Interspeech.2007-345

Abstract

Previous work on bandpass modulation filtering for noise suppression has resulted in unwanted perceptual artifacts and decreased speech clarity. Artifacts are introduced mainly due to half-wave rectification, which is employed to correct for negative power spectral values resultant from the filtering process. In this paper, modulation frequency estimation (i.e., bandwidth extension) is used to improve perceptual quality. Experiments demonstrate that speech-component lowpass modulation content can be reliably estimated from bandpass modulation content of speech-plus-noise components. Subjective listening tests corroborate that improved quality is attained when the removed speech lowpass modulation content is compensated for by the estimate.

Index Terms: Modulation filtering, noise suppression, Hilbert envelope, perceptual quality, rectification.

1. Introduction

With the advances in speech communication technologies, noise suppression (NS) has become an integral component in applications such as hearing aids, mobile phones, and voice controlled systems. Commonly, with single-microphone NS algorithms, voice activity detection (VAD) is employed to estimate (update) noise statistics during speech pauses. In practice, however, VAD performance degrades substantially for low signal-to-noise ratio (SNR), often leading to unreliable speech pause decisions and poor noise suppression performance. As a consequence, single-microphone NS algorithms that do not depend on VAD are highly desirable. One such approach consists of filtering the temporal trajectories of the short-time spectrum of speech. This approach is commonly referred to as "modulation filtering" and spectral content (usually due to noise) which change slower or faster than the typical range of change of speech are removed.

In [1], the importance of different modulation frequencies in speech intelligibility was investigated. The speech signal was split in different frequency bands and the spectral magnitude envelope of each band was lowpass filtered by different modulation filters. It is reported that modulation frequencies below 16 Hz play an important role in speech intelligibility. In [2], a similar experiment is carried out with highpass modulation filters. In this second study, modulation frequencies above 4 Hz are shown to be important. Tests with bandpass modulation filters are described in [3] and frequencies between 1 Hz and 16 Hz are shown to be important.

Modulation filtering is used by the speech recognition community to improve recognition accuracy under noisy conditions (e.g., [4]). For recognition, bandpass modulation filtering serves two roles: removal of modulation content not consistent with normative speech behavior, and suppression of the long-term spectrum of the speech signal. The latter has been shown to improve speaker independence [5]. In [4], RASTA (RelAtive SpecTrA) filters are applied for noise suppression; colored musical noise and little improvement in speech intelligibility was reported. In [6], the design of Wiener-like modulation filters from clean and noisy training data is described. The filters are shown to be mostly efficient on disturbances similar to those present in the training data. An adaptive model is then proposed where pre-designed modulation filters are chosen based on an online estimated SNR [7]. Perceptual artifacts, such as noise with periodic level fluctuations, were reported in scenarios where the test conditions were similar to those present in training.

The removal of lowpass modulation spectral content is not beneficial for NS applications as the resultant power spectra may assume negative values. As with the spectral subtraction literature, a half-wave rectifier is commonly used. Rectification, however, may result in unwanted perceptual artifacts and in reduced speech clarity. The approach used in [4, 6, 7] to reduce such artifacts was to filter the cubic-root compressed power spectrum.

In this paper, an alternate route to avoid rectification artifacts is taken; namely, bandwidth extension is performed. We propose to estimate the "speech-only" low frequency modulation content from the speech-dominated bandpass modulation content of the noisy signal. Our experiments show that the Hilbert envelope of temporal trajectories of the bandpass filtered noisy signal can serve as a reliable estimator of the lowpass modulation content of clean speech. Subjective tests are carried out and perceptual artifacts that arise from bandpass modulation filtering are shown to be greatly reduced once the removed lowpass modulation content is accounted for by the estimate.

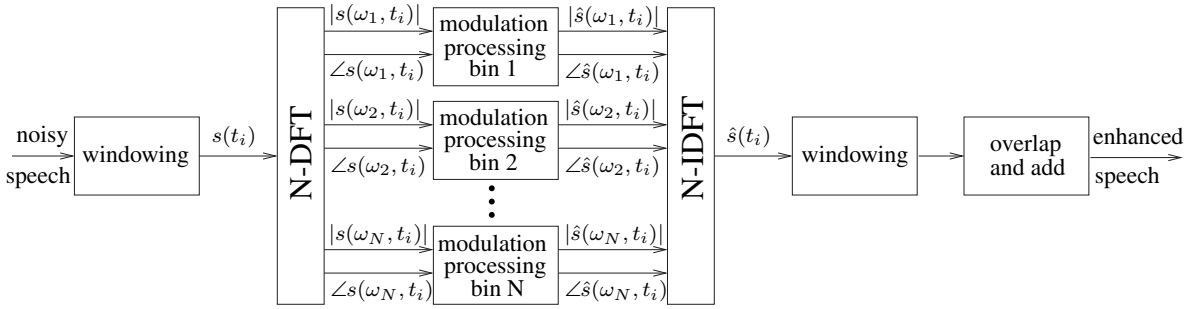

Figure 1: Block diagram of the proposed "compensated modulation filtering" method.

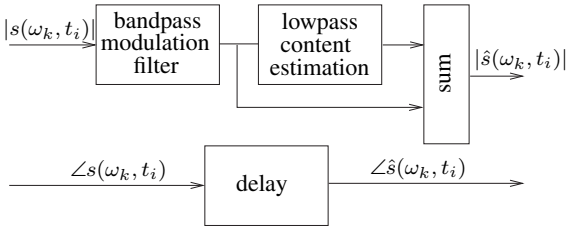

Figure 2: Diagram of k^{th} modulation processing module.

2. Modulation Filtering: Methodology

In this paper, modulation filtering is described as filtering of temporal trajectories of a short-term spectral component. For the sake of notation, let $s(\omega_k, t_i), k = 1, \ldots, N$ and $i = 1, \ldots, T$, denote the short-term spectral component at the k^{th} frequency band and i^{th} time step of the short-term analysis. N and T denote total number of frequency bands and time steps, respectively. For a fixed frequency band ω_k, $s(\omega_k, t_i), i = 1, \ldots, T$, represents the band temporal trajectory. Next, we describe the signal processing involved in our modulation filtering experiments and the proposed enhancements.

2.1. Signal Processing

We use the Gabor transform for spectral analysis. The Gabor transform is a unitary transform (energy is preserved) and consists of an inner product with basis functions that are windowed complex exponentials. In our experiments, doubly over-sampled Gabor transforms are used and implemented based on discrete Fourier transforms (DFT), as depicted in Fig. 1. First, the noisy signal is windowed by a power complementary window (more details in Section 3.1). An N-point DFT is then taken and the magnitude ($|s(\omega_k, t)|$) and phase ($\angle s(\omega_k, t)$) components of each frequency bin are fed into what we call a "modulation processing" module. Bandpass modulation filtering and the proposed enhancements are carried out in this module, as depicted in Fig. 2.

The magnitude trajectory $|s(\omega_k, t)|$ is bandpass filtered with a linear phase FIR filter. If only bandpass modulation filtering is being performed, half-wave rectification would occur in the sequel. Here, to avoid artifacts resultant from rectification, an estimate of the removed lowpass modulation content of the speech component is used. With noise corrupted speech, the lowpass modulation content carries information of both speech *and* noise. For stationary noise, in particular at low SNRs, the noise component of the lowpass modulation content tends to dominate the speech component. Conversely, information obtained from modulation frequencies between 1-16 Hz are attributed mainly to speech [3]. As a consequence, we propose to estimate the speech lowpass modulation content from the speech-dominated bandpass modulation content of the noisy signal, as depicted in Fig. 2.

The remaining modulation processing step consists of delaying the phase by an integer number of samples; the delay is dependent on the order of the filter used. The outputs of the k^{th} modulation processing module are the estimated lowpass modulation content combined with the bandpass modulation content ($|\hat{s}(\omega_k, t_i)|$) and the delayed phase components ($\angle\hat{s}(\omega_k, t_i)$). An N-point IDFT is taken and the modified signal is again windowed by the power complementary window. Overlap-and-add is used to reconstruct the enhanced signal. We refer to the overall proposed method as "compensated modulation filtering."

2.2. Importance of Lowpass Modulation Content

As mentioned previously, negative power spectral values may result from bandpass modulation filtering. We argue that if bandwidth extension is performed, the number of such "negative-spectra" instances is reduced and improved quality is attained. To validate this claim, an experiment is carried out where the "lowpass estimation content" block in Fig. 2 is replaced by the true clean speech lowpass modulation content. In this pilot experiment, we also examine two linear phase FIR complementary lowpass and bandpass filters; the filter magnitude responses are depicted in Fig. 3. The first pair (solid) consists of a lowpass filter with cutoff frequency at 4Hz and a bandpass filter with cutoff frequencies at 4 Hz and 16 Hz, following [1, 2]. The second filter pair (dotted) follows [3] and has cutoff frequencies at 1 Hz and 16Hz.

As expected, informal listening tests carried out in

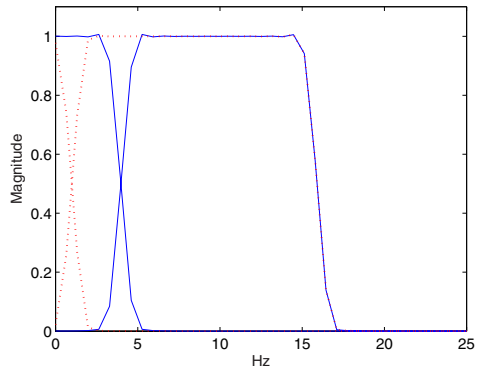

Figure 3: Magnitude response of two pairs of complementary low and bandpass modulation filters. First pair (solid) has cutoff frequencies at 4Hz and 16Hz. The second pair (dotted) has cutoff frequencies at 1Hz and 16Hz.

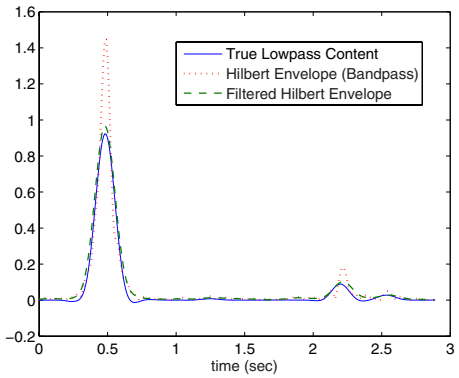

Figure 4: Modulation Hilbert envelopes of bandpass content (dotted), lowpass filtered Hilbert envelope (dashed), and true lowpass modulation content (solid). Plots are for the sixth frequency bin, corresponding to 300Hz.

our labs show that when lowpass information is not compensated for, perceptual artifacts due to half-wave rectification are more pronounced with the bandpass modulation filter with lower cutoff frequency at 4 Hz. As an example, for the 2.5 second noisy signal depicted in Fig. 5, such filter results in a half-wave rectification activation rate of 0.304; the activation rate for the filter with lower cutoff frequency at 1 Hz is 0.256. Once lowpass modulation content is compensated for, artifacts are greatly reduced and speech clarity is increased. Rectification activation rates were reduced to negligible values for both filter pairs. The enhanced quality obtained from this experiment serves as an upper bound on the attainable quality of modulation filtering and emphasizes the need for an effective estimator of the lowpass modulation content.

2.3. Improved-Quality Modulation Filtering

In order to reduce perceptual artifacts resultant from half-wave rectification, we propose the use of the Hilbert envelope of temporal trajectories of the bandpass filtered noisy signal as an estimate of the speech lowpass modulation content. The Hilbert envelope (henceforth referred to as modulation Hilbert envelope) gives a measure of the instantaneous energy as a function of time and, in our experiments, has shown to be highly correlated with the lowpass modulation content of clean speech. The plots in Fig. 4 assist in illustrating this behavior.

As can be seen, the modulation Hilbert envelope of the bandpass content (dotted) is highly correlated with the true lowpass modulation content (solid). A better estimate is attained if the modulation Hilbert envelope is projected onto the space of the removed lowpass modulation content; this can be attained by filtering the envelope with the complementary lowpass filter. As seen from Fig. 4, the filtered envelope (dashed) is a closer match to the true lowpass modulation content. Moreover, lowpass filtering of the envelope is important, in particular for higher frequency bins, where noise components tend to dominate speech components. The waveforms in Fig. 5 illustrate the noise reduction capabilities of the proposed method (BP+Hilb). For comparison purposes, waveforms of the signals processed by bandpass modulation filtering (BP), the hypothetical scenario where the true lowpass content is used (True), and EVRC are also illustrated. The original signal is uttered by a male speaker and is corrupted by white noise at SNR$=5$ dB.

3. Experimental Results

In this section, implementation details and two subjective listening tests which demonstrate the effectiveness of the proposed Hilbert envelope-based estimator are described.

3.1. Implementation Details

In our experiments, a square-root Hann window of length 20 milliseconds with 50% overlap and a frame period of 10 milliseconds is used for the Gabor transform. In order to attain accurate resolution at 1 Hz, higher order filters are needed. Here, 131- and 51-tap linear phase filters are used to implement bandpass filters with lower cutoff frequencies at 1 Hz and 4 Hz, respectively. As mentioned in Section 2.2, perceptual artifacts resultant from half-wave rectification are less pronounced when the filter with lower cutoff frequency at 1 Hz is used. In order to perform fair subjective tests, as described below, a filter with such cutoff frequency is used in the tests. We emphasize, however, that on our databases, the proposed compensated modulation filtering method resulted in similar quality with either filter. As an example, a rectification activation rate of 0.05 is attained for both filters; this compares favorably with the rates reported in Section 2.2. As a consequence, significant decrease in delay can be attained with the proposed method if the bandpass filter with lower cutoff frequency at 4Hz is used.

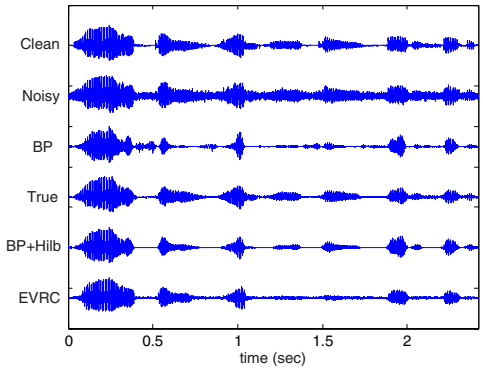

Figure 5: Waveforms, top to bottom: clean, noisy, (rectified) bandpass filtering, bandpass with true lowpass content, compensated modulation filtering, and EVRC NS.

3.2. Subjective Listening Tests

Two "original–A–B" listening tests were performed with five expert listeners. The first test (T1) served to compare the quality of (rectified) bandpass modulation filtering with and without lowpass content compensation. The second test (T2) served to compare the quality of the proposed scheme with EVRC-NS [8]. In each test, listeners were presented with 24 sentences uttered by four speakers (2 male and 2 female). The speech was corrupted by three noise types (babble, plane, and white noise) at two SNR levels (5 dB and 10 dB) and processed by the different NS algorithms. For each of the 24 utterances, the listeners were presented the original (clean) utterance first, followed by, in random order, noise suppressed utterances produced by the two NS algorithms being tested. To avoid a bias based on the presentation order, at a random point throughout the test the listeners were presented the noise suppressed utterances again, but in reversed order. Before making a final judgement, the subjects were allowed to listen to the same sequence as often as needed. Listeners were asked to judge whether excerpt A or B sounded closer to the original.

The results in Table 1 show the percentage preference of the proposed compensated modulation filtering approach over bandpass modulation filtering (T1) and over EVRC (T2). As can be seen, tests T1 and T2 rendered, on average, 86.25% and 70.85% preference for the compensated modulation filtering scheme, respectively. Somewhat higher improvement is attained for stationary white noise and plane engine noise. Note that at low SNR, the EVRC algorithm limits its noise attenuation to avoid severe signal distortions. Informal conversations with the listeners suggested that at times, the EVRC residual noise was preferred over the artifacts generated by the compensated filtering scheme. On the other hand, some listeners were displeased with the fact that some phonemes were completely attenuated with EVRC noise suppression. The plot in Fig. 5 illustrates one such example. Excerpts from a subset of the audio samples provided to the listeners has been made available online [9].

Table 1: Preference (in terms of percentage) of the proposed compensated modulation filtering approach over bandpass modulation filtering (T1) and EVRC-NS (T2).

Subjective Test	Noise Type		
	Babble (%)	Plane (%)	White (%)
T1	85.00	86.25	87.50
T2	68.75	70.00	73.75

4. Conclusion

We have shown that the quality of bandpass modulation filtered noisy speech can be enhanced if lowpass modulation content is estimated from speech-dominated bandpass modulation content. One estimator, based on Hilbert envelopes of bandpass filtered temporal trajectories, is proposed and subjective listening tests corroborate that improved quality is attained. We emphasize that, other than the fact that normal speech lies in the 1-16 Hz modulation frequency range, the proposed method does not take into account any other knowledge of properties of speech nor of noise statistics. Alternate estimators are currently being investigated.

5. References

[1] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Ac. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[2] ——, "Effect of reducing slow temporal modulations on speech reception," *J. Ac. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[3] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. Intl. Conf. Speech and Lang. Proc.*, Oct. 1996, pp. 2490–2493.

[4] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 4, Oct. 1994.

[5] C. Nadeu, P. Paches-Leal, and B.-H. Juang, "Filtering the time sequences of spectral parameters for speech recognition," *Speech Commun.*, vol. 22, 1997.

[6] H. Hermansky, E. Wan, and C. Avendano, "Speech enhancement based on temporal processing," in *Proc. Intl. Conf. Audio, Speech, and Signal Proc.*, 1995.

[7] C. Avendano, H. Hermansky, M. Vis, and A. Bayya, "Adaptive speech enhancement using frequency-specific SNR estimates," in *Proc. Int. Voice Tech. Telecom. Appl. (IVTTA)*, Oct. 1996, pp. 65–68.

[8] 3GPP2 C.S0014-0, "Enhanced variable rate codec (EVRC)," Dec. 1999.

[9] http://qshare.queensu.ca/Users01/2thf/www/index.htm