Citation for the published Conference paper:

Title:
Modulation Domain Adaptive Gain Equalizer for Speech Enhancement

Author:
Muhammad Shahid, Rizwan Ishaq, Benny Sällberg, Nedelko Grbic, Benny Lövström, Ingvar Claesson

Conference Name:
The IASTED International Conference on Signal and Image Processing and Applications

Conference Year:
2011

Conference Location:
Crete, Greece

MODULATION DOMAIN ADAPTIVE GAIN EQUALIZER FOR SPEECH ENHANCEMENT

Muhammad Shahid, Rizwan Ishaq, Benny Sällberg, Nedelko Grbic, Benny Lövström and Ingvar Claesson
Department of Signal Processing, Blekinge Institute of Technology, SE-37179 Karlskrona, Sweden
Correspondence author: muhammad.shahid@bth.se

ABSTRACT
This paper evaluates speech enhancement by filtering in the modulation frequency domain, as an alternative to filtering in conventional frequency domain. Adaptive Gain Equalizer (AGE) is a commonly used single-channel speech enhancement algorithm. A recently introduced class of signal transformations called modulation transform has successfully made its place alongside classical time/frequency representations. This paper presents an implementation of AGE within modulation system, for the purpose of enhancing the speech signal. The successful implementation of the proposed system has been validated with various performance measurements, i.e., Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). A spectrogram analysis is also presented to further substantiate the performance of this work.

KEY WORDS
Speech enhancement, Adaptive gain equalizer, Modulation domain.

1 Introduction

Speech as the main part of the communication systems, is usually degraded during the transmission by different types of noise, e.g., Gaussian noise, engine noise, periodic noise and other interferences. There are a variety of methods for reduction of noise from speech signal, e.g., spectral subtraction (frequently used for noise reduction) [1] and optimum Wiener filtering [2]. Adaptive Gain Equalizer (AGE) [3] is a noise reduction method that focuses on enhancing the speech signal instead of suppressing the noise. The speech enhancement is carried out by weighting the subbands in time-frequency domain according to an estimate of the Signal-to-Noise Ratio (SNR). This method offers better result in terms of low complexity, low delay, low distortion and there is no need for Voice Activity Detector (VAD) .

The modulation system assumes that a speech signal is composed of a modulator and a carrier. The signal is represented by,

$$x(t) = m(t)c(t) \tag{1}$$

where $m(t)$ denotes the low frequency part of the signal, called modulator, and it modulates a high frequency carrier $c(t)$. Studies have shown that the modulators of speech signal are most important for the intelligibility of the speech signal. The importance of the modulator in speech signals brought the attention of many researchers .

AGE implementation has been intended so far in time-frequency domain, but here an implementation of AGE in a modulation system is proposed. Modulation systems which are based on sub-band modulators, perfectly fit the AGE system which works on the sub-bands of the signal.

1.1 Literature Survey

Zadeh [4] is considered to be the pioneer of the field of modulation domain who suggested a two dimensional bi-frequency system, where time variation of the acoustic frequency is the second dimension of frequency. Atlas et al. used the concept of coherent modulation for the target talker enhancement in speech enhancement [5]. They proved that modulation domain moderately increases the speech intelligibility . Coherent modulation using the frequency reassignment has been used for speech enhancement and for demodulation of a signal into modulator and carrier [6]. Li et al. described the theory behind modulation filtering which offers a new approach to modifying non-stationary signals e.g., speech. They presented the coherent modulation analysis based on instantaneous frequency estimation using conditional mean frequency. In addition, they showed that the proposed method accurately estimates the carriers and modulators of the signals [7]. Speech polluted by wind noise has been enhanced by using coherent modulation comb filtering by King et al. [8]. Although the modulation filtering has mostly been used for the purpose of speech enhancement, Vinton et al. also used it for audio compression. They showed that a 32 kb/s/channel outperformed MPEG-1 coded at 56 kb/s/channel (both at 44.1 kHz), using the modulation technique [9]. The concept of homomorphic demultiplication is connected to the modulation spectral analysis/synthesis and it was outlined by Atlas et al. in [10]. Clark et al. showed in [11] the effectiveness of modulation filtering by measuring the empirical modulation frequency response and got a near-ideal response performance, and 25 dB improvement has been shown for suppressing undesired modulation frequencies over incoherent modulation. Clark presented the Center of Gravity (COG) method for decomposition of a sub-band signal, and he used coherent modulation filtering for the interpolation of long gaps in acoustic signals [12].

The concept of AGE for the reduction of noise in speech signals, has shown its success in real time and proven to be a low complexity system [3]. The method used an FIR filter bank to get the required results and it was also shown that the system adapted itself for different types of noise. The proposed AGE method using the mixed analog and digital hybrid approach yielded around 13 dB speech enhancement [13]. The AGE was originally intended for the digital domain, but [14] provides an analog implementation which does not use quantization and digitization and it is also best fitted for battery powered applications. A hybrid solution to overcome problems related to a digital and an analog implementation of the AGE is found in [15].

1.2 Main Contribution

The main contribution of this paper is to combine the AGE and modulation system domain for speech enhancement. Hence, the advantage of benefits from both of the fields has been taken to build up a new system. This approach has proven to be robust, flexible in implementation and has been validated by performance measures like Signal to Noise Ratio Improvement (SNRI), Mean Opinion Score (MOS) and Spectral Distortion (SD). Section 2 briefly introduces the modulation system, section 3 introduces the concept of AGE and its operation in the modulation frequency domain and section 4 evaluates the proposed system. Section 5 concludes this work with a summary and future research directions in the area.

2 Modulation System

A modulation domain spectrum is obtained from a certain acoustic spectrum by taking short-time Fourier transform (STFT) of the speech signal at the given acoustic frequency. The speech signal modulators are the most important components for speech intelligibility. Shamma [16] reported that auditory cortex neurons possibly decompose the acoustic contents into spectro-temporal modulation contents. It has been found that if the modulators of the speech signal are replaced by constant amplitude modulators, while carriers are preserved, speech is not intelligible. However when the modulators are preserved but carriers are altered, the speech is intelligible [17]. Modulation domain actually decomposes the speech, or other natural signals, into modulators and carriers whereafter the modulators of the signals are analyzed. A general framework for modulation frequency domain analysis, and filtering is given in figure 1. A modulation frequency system is described by the following steps:

- Filter bank to get sub-band signals
- Demodulation i.e., decomposition of each sub-band signal into a modulator and a carrier.

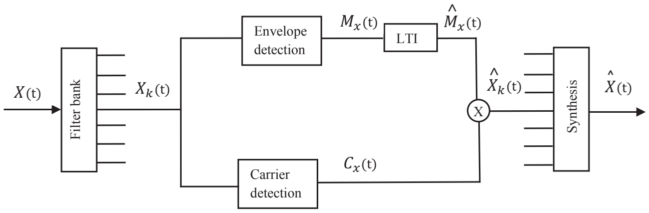

Figure 1. A general framework of the modulation filtering and analysis system [17]

- Analysis of the modulators of the sub-band signals by discrete Fourier transform of each modulators
- Modification of the modulators (e.g. linear filtering)
- Re-modulation (recombination of modified modulators with original carriers)
- Synthesis of signals

The modulation system filter bank divides the wide-band signal into K narrow-band sub-bands. The signal $x(t)$ is passed through the filter bank's set of band-pass filters h_k, which renders the sub-band signals $x_k(t)$.

$$x_k(t) = h_k * x(t) \tag{2}$$

where $*$ denotes the convolution operator. The demodulation process decomposes the sub-band signal into its envelope and carrier. Its efficient to decimate the sub-band signals so that the redundant samples may be removed. Modification of the modulators is done by the modulation filtering which mostly uses linear time invariant filters $g(t)$, i.e., $\hat{m_k}(t) = m_k(t)g(t)$. A modulation spectrogram and modulation analysis can be done by computing the Fourier transform along the time-axis of the spectrogram (magnitude) or by utilizing the spectrum of the envelop signals, which gives the modulation frequency along horizontal axis and acoustic frequency along vertical axis. Re-modulation is the process in which modified modulators $\hat{m_k}(t)$ are combined with the original carriers, obtained in the process of demodulation, to get the modified sub-band signals $\hat{x}_k(t)$. The synthesis process reconstructs the modified signal $\hat{x}(t)$ using the modified sub-band signals $\hat{x}_k(t)$, according to the following equation. Interpolation must be performed prior to this stage if decimation was done before.

$$\hat{x}(t) = \sum_{k=1}^{K} \hat{x}_k(t) \tag{3}$$

Envelope detection is used for demodulation of a signal and it is the most important part of the modulation frequency system. There are two types of envelope detectors mostly used, coherent envelope detection and incoherent envelope detection. Magnitude, or magnitude-like, operations are used to estimate modulators in incoherent detection, while coherent detection use the carrier estimate operations. Incoherent envelope detection detects the envelope and carrier independently and coherent detection uses the carrier estimation for the calculation of the envelope. Following is a brief description about one of the methods used for coherent carrier detection which is used in this work.

2.1 Spectral Center of Gravity Carrier Estimation

In this recently introduced method of the center-of-gravity approach, instantaneous frequency $\omega_k(n)$ is defined as instantaneous spectrum average frequency of $x_k(t)$ at time t [18]. An instantaneous spectrum with short-time Fourier transform is computed as,

$$S_k(\omega, t) = \sum_p g(p) x_k(t+p) e^{-j\omega p} \tag{4}$$

where $g(p)$ is a short spectral-estimation window. The instantaneous frequency $\omega_k(t)$ of the sub-band signal $x_k(t)$ is estimated as,

$$\omega_k(t) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega,t)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega,t)|^2 d\omega} \tag{5}$$

The phase $\phi_k(t)$ of the carrier is computed as follows

$$\phi_k(t) = \sum_{p=0}^{t} \omega_k(p) \tag{6}$$

The carrier c_k is

$$c_k(t) = e^{j\phi_k(t)} \tag{7}$$

and the complex valued modulator $m_k(t)$ is given by

$$m_k(t) = x_k(t) c_k^*(t) \tag{8}$$

3 Adaptive Gain Equalizer System

As discussed in [3], the AGE consists of a filter bank with different band-pass filters. Each sub-band is weighted by a gain function which amplifies the signal when speech is present and keeps the noisy part of the signal, where no speech is present, to unity. A filter bank of K bandpass filters divides the input signal $x(n)$ into K sub-bands $x_k(n)$.

$$x_k(n) = h_k * x(n) \tag{9}$$

Here h_k is the impulse response of the filter bank sub-band k and $*$ denotes the convolution. The time domain signal is modeled as a sum of sub-band signals, according to:

$$x(n) = \sum_{k=1}^{K} x_k(n) = \sum_{k=1}^{K} (s_k(n) + w_k(n)) \tag{10}$$

where $s_k(n)$ is the desired speech signal related to k^{th} sub-band, while $w_k(n)$ is the additive noise in the sub-band k.

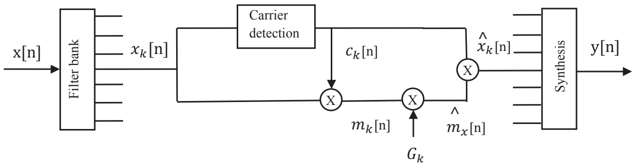

Figure 2. Adaptive gain equalizer in modulation domain

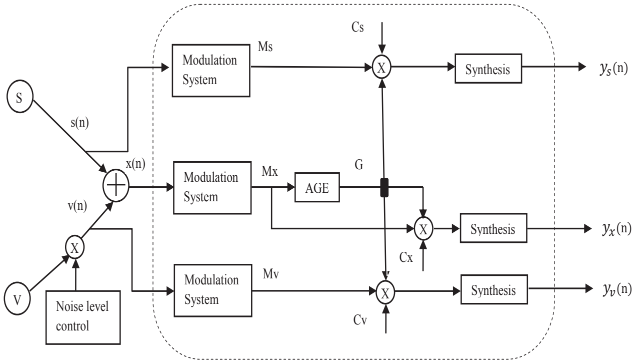

Figure 3. Experiment setup

The output signal $y(t)$, with the amplified speech signal, is computed as

$$y(n) = \sum_{k=1}^{K} G_k(n) x_k(n) \tag{11}$$

where $G_k(n)$ is the AGE weighting function which amplifies the signal when speech is active.

3.1 Gain Function

Two terms used for the calculation of the gain function are; a long term (slow) average $A_{s,k}(t)$ and the short term (fast) average $A_{f,k}(t)$. The short term average, for sub-band k, $A_{f,k}(n)$ is calculated as

$$A_{f,k}(n) = \alpha_k A_{f,k}(n-1) + (1-\alpha_k) \mid x_k(n) \mid \tag{12}$$

where α_k is a small positive constant, given by

$$\alpha_k = \frac{1}{T_{s,k} F_s} \tag{13}$$

where F_s is the sampling frequency in Hz and $T_{s,k}$ is a time constant in seconds. In the same manner, a slow average is computed as

$$A_{s,k}(n) = (1+\beta_k) A_{s,k}(n-1) \tag{14}$$

if $A_{s,k}(n-1) \leq A_{f,k}(n)$, and

$$A_{s,k}(n) = A_{f,k}(n) \tag{15}$$

if $A_{s,k}(n-1) > A_{f,k}(n)$
where β_k is a small positive constant. The AGE gain function is computed as:

$$G_k(n) = \left(\frac{A_{f,k}(n)}{A_{s,k}(n)} \right)^{p_k} \quad (16)$$

where $p_k \geq 0$, and $A_{s,k}(n) > 0$.

3.2 Modulation Domain AGE

The functionality of the AGE has been extended to work in the modulation domain for speech enhancement. Modulation domain separates each sub-band signal into a carrier and a modulator. While only modulators are considered here, the AGE is implemented on each modulator to enhance the speech. The system is shown in figure 2. The mathematics for AGE in the modulation domain is the same as for AGE in the sub-band domain, the long term average and the short term average are calculated for each sub-band modulator, instead of the sub-band itself. The gain function is multiplied with the modulator of the sub-band to yield a modified modulator $\hat{m}_k(n)$ which is then used with the carrier in the reconstruction stage of the modulation system.

$$\hat{m}_k(n) = m_k(n)G_k \quad (17)$$

$$\hat{x}_k(t) = c_k(n)\hat{m}_k(n) \quad (18)$$

The synthesized signal $y(n)$ is finally calculated by adding up all the components.

$$y(n) = \sum_{k=1}^{K} \hat{x_k}(n). \quad (19)$$

The gain function G_k is given by

$$G_k = \min\,(L, \frac{A_{f,k}}{L_{opt}.A_{s,k} + \epsilon}) \quad (20)$$

where $A_{f,k}$ denotes short term average and $A_{s,k}$ denotes the long term average, L is a limiting threshold which limits the gain function's value and L_{opt} is an optimum level of control on the value of the gain function. The averages are computed as:

$$A_{f,k}(n) = \alpha_f A_{f,k}(n-1) + (1-\alpha_f) \mid m(n) \mid \quad (21)$$

$$A_{s,k}(n) = \alpha_s A_{s,k}(n-1) + (1-\alpha_s) \mid m(n) \mid \quad (22)$$

$$A_{s,k}(n) = \min\,(A_{s,k}(n), A_{f,k}(n)) \quad (23)$$

where α_f and α_s are time constants of the short term and long term averages, respectively.

4 Evaluation of The Proposed System

Figure 3 shows the experimental setup, where $s(n)$ is the clean speech signal, $v(n)$ is a noise signal and $x(n)$ is the sum of speech and noise signals $(s(n) + 10^{\frac{-SNR}{20}} v(n))$ scaled by desired level of Signal to Noise Ratio (SNR). M_s, C_s, M_x, C_x, M_v and C_v are the signal matrices of modulators and carriers for $s(n)$, $x(n)$ and $v(n)$ respectively. The gain matrix G is calculated by passing M_x through AGE system. This G is then multiplied with the M_x, M_s and M_v, whereafter the re-modulation and the synthesis processes generate the output signals $y_x(n)$, $y_s(n)$, $y_v(n)$, as depicted in figure 3. The system was evaluated with the following parameter setings. $L = 1$, L_{opt}= 1 to 20, $T_s = 4s$ and $T_f = 0.04s$. The speech signals comprise male F_s=16 kHz and female F_s=16 kHz speech signals and the noise signals are scaled so as to have 10 dB, 5 dB, 0 dB and -5 dB SNR. Noise signals used were Engine Noise (EN), Factory Noise (FN), Gaussian Noise (GN), Tonal Noise (TN) and Impulse Noise (IN). The performance measurement was evaluated by the Signal to Noise Ratio Improvement (SNRI), Perceptual Evaluation of Speech Quality (PESQ) and Spectral Distortion (SD). SNRI of male speech signal for TN at 0 dB SNR with L_{opt} = 20 was around 10 dB and for other noises was between 4 dB to 6

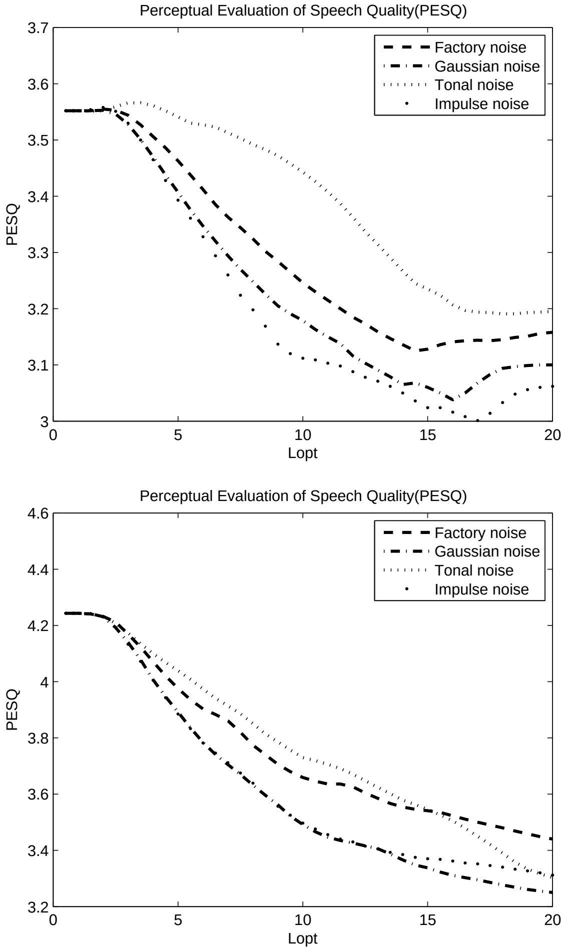

Figure 4. MOS for the processed male speech signal (upper) and female speech signal (lower) with noise at 10 dB SNR

Table 1. Spectral distortion (SD) results

Noise SNR	0 dB		10 dB	
L_{opt} **range**	0 to 5	5 to 20	0 to 5	5 to 20
Speaker	Male			
SD for FN [dB]	-18 to -4	-4 to -2	-18 to -4	-4 to -2
SD for IN [dB]	-18 to -4	-4 to -2	-18 to -4	-4 to -2
SD for TN [dB]	-18 to -6	-4 to -2	-18 to -4	-6 to -2
Speaker	Female			
SD for FN [dB]	-34 to -12	-12 to -2	-34 to -15	-15 to -4
SD for IN [dB]	-34 to -15	-15 to 0	-18 to -6	-6 to -2
SD for TN [dB]	-34 to -15	-15 to -7	-34 to -18	-18 to -8

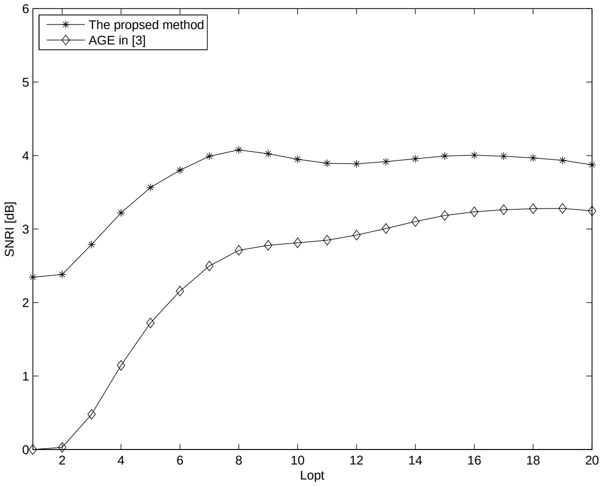

Figure 5. SNRI plots of two speech enhancement methods

dB. The female speech signal also had SNRI of 9 dB for TN and around 3 to 5 dB for EN, FN, GN, IN at 0d B SNR. PESQ has been calculated by comparing $s(n)$ and $y_s(n)$ which gives an objective measure of how much degradation the system has introduced on the speech signal due to introducing the AGE gain function. The objective Mean Opinion Score (MOS) as computed by the PESQ for most of the tests given above was 3, which is considered fair for speech signals. Experiments have been performed to find out the optimal value on the critical system parameter L_{opt}, for different noise cases and for different speaker situations. Figure 4 shows the MOS values for both male and female speech signals at 10 dB of noise SNR. It is interesting to note that female speech has higher values of MOS than male speech under similar conditions. This observation is attributed to the fact that female speech with higher pitch is less affected by some noises. Moreover, the SD is very low for $L_{opt} < 5$ and then increases rapidly with increasing L_{opt} values for all tests. For male speech signal, the SD at L_{opt} = 20 is around -2 dB and -4 dB for FN,GN,TN and IN and some of them are shown in table 1. The female speech signal has different behavior than the male speech signal on SD. For female speech, SD is found to be -2 dB for EN, GN, IN and -4 dB for FN and -8 dB for TN at the L_{opt}=20.

The proposed method was also compared against the speech enhancement method by AGE proposed in [3]. It was observed that the proposed method has better performance than the reference method of [3]. One such comparison is shown in figure 5 where a male speech signal having mixed with 5 dB SNR factory noise is enhanced by two methods and the proposed method clearly outperforms its counterpart in [3].

4.1 Spectrogram Analysis

The spectrogram of a male speech signal that has been mixed with gaussian noise at 10 dB SNR and the spectrogram after enhancement with the propsed AGE system, are given in figure 6. The AGE algorithm converges after 0.2 seconds for all test cases, whereafter it may be observed that the disturbing noise is reduced while the formants of the speech are maintained. Enhanced signal $y_x(n)$ has shown the formants very clearly after the processing. Although the Gaussian noise is spread throughout the frequency plane, the AGE works very efficiently, but a little bit speech signal energy has also been lost. The spectrogram of male speech signal mixed with tonal noise at 0 dB SNR and enhanced male speech signal by the AGE was also observed. The tonal noise which had all of its energy around 1 kHz has been reduced by the AGE, i.e., reduced its energy, while maintaining the formants of speech. Moreover, the impulse noise at 0 dB SNR, which is similar to gaussian noise in spreading its energy through all the frequencies, has been successfully eliminated.

5 Conclusion

A novel approach of speech enhancement in modulation frequency domain has been explored and the promising results obtained by using the proposed method have been presented in this paper. The adaptive gain equalizer (AGE), which has shown its advantages already in digital, analog

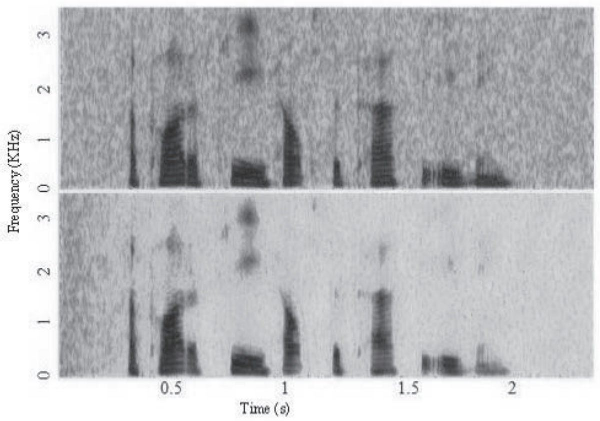

Figure 6. Spectrogram of noisy male speech (upper) having Gaussian noise at 10 dB SNR and the enhanced signal by the proposed method (lower)

and hybrid domains by its simplicity, low complexity for being robust to different noisy environments, has been implemented in the modulation frequency domain in this paper. The detailed analysis of the system has put light on its advantages and disadvantages, i.e. where the evaluation section highlights the compromise between low SD and high SNRI. The system provides good improvement on the female speech signal, with better SNRI, low SD, fair MOS, and output speech signal sounds good. The maximum SNRI obtained for the female speech signal analysis was approximately 9 dB and SD of female speech for some noise has been shown 0 dB.

The spectrogram analysis provides another view of these results. The AGE gain function adapts during the first 0.2 seconds. This start-up time can be reduced by varying the integration time, but changing the integration time has obvious consequences on the signal integrity and the noise reduction performance. Moreover, the proposed method has shown its potential as a better alternative to the traditional methods of speech enhancement.

Future work is to implement this system in real time and other speech enhancement methods may also be tried in modulation domain.

References

[1] S. F. Boll, Suppression of acoustic noise in speech using spectral subtraction, *IEEE trans. Accoust. Speech and Sig. Proc.*, 27(2), 1979, 113-120.

[2] M. H. Hayes, *Statistical Digital Signal Processing and Modeling* (New York: John Wiley and Sons Inc., 1996).

[3] N. Westerlund, M. Dahl and I. Claesson, Real-time implementation of an adaptive gain equalizer for speech enhancement purposes, *Proc. 2nd WSEAS International Conf. on Electronics, Control and Signal Processing*, Singapore, 2003, 2:1–2:8.

[4] L. Zadeh, Frequency analysis of variable networks, *Proc. IRE*, 38(3), 1950, 291-299.

[5] L. E. Atlas and S. M. Schimmel, Target talker enhancement in hearing devices, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, 4201-4204.

[6] S. M. Schimmel, K. R. Fitz and L. E. Atlas, Frequency reassignment for coherent modulation filtering , *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, 261-264.

[7] Q. Li and L. Atlas, Coherent modulation filtering for speech, *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, 4481-4484.

[8] B. King and L. Atlas, Coherent modulation comb filtering for enhancing speech in wind noise, *Proc. International Workshop on Acoustice Echo and Noise Control*, Seattle, USA, 2008.

[9] M. S. Vinton and L.E. Atlas, A scalable and progressive audio codec, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, USA, 2001, 3277-3280.

[10] L. Atlas, Q. Li and J. Thompson, Homomorphic modulation spectra, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, 2004, 761-764.

[11] C. P. Clark and L. Atlas, A sum-of-product model for effective coherent modulation filtering, *Proc. IEEE, International Conference on Acoustics, Speech and Signal Processing*, Taipei , China, 2009, 4485-4488.

[12] C. P. Clark, Effective coherent modulation filtering and interpolation of long gaps in acoustic signals, *Master thesis*, University of Washington, 2008.

[13] M. Dahl and I. Claesson and B. Sällberg and H. Akesson, A mixed analog-digital hybrid for speech enhancement purposes, *Proc. IEEE International Symposium on Circuits and Systems*, Kobe , Japan, 2005, 852- 855.

[14] M. Dahl and I. Claesson and B. Sällberg and H. Akesson, A mixed analog-digital hybrid for speech enhancement purposes, *Proc. IEEE International Symposium on Circuits and Systems*, Kobe, Japan, 2005, 852- 855.

[15] B. Sällberg, M. Dahl, Speech Enhancement implementations in the digital, analog and hybrid domain, *Proc. Swedish System on Chip Conference*, Stockholm, Sweden, 2005.

[16] S. Shamma, Encoding sound timbre in the auditory system, *IETE Journal of research*, 49(2), 2003, 193-205.

[17] S.M. Schimmel, Theory of modulation frequency analysis with applications to hearing devices, *Ph.D. dissertation*, University of Washington, 2007.

[18] P. Clark and L. E. Atlas, Time-frequency coherent modulation filtering of non-stationary signals, *IEEE transaction on Signal Processing*, 57(11), 2009, 4323-4332.