Subband Modulator Kalman Filtering for Single Channel Speech Enhancement

Rizwan Ishaq*, Begoña García Zapirain*, Muhammad Shahid** and Benny Lövström**

*University of Deusto, Bilbao, Spain
**School. of Elec. Engineering, Blekinge Institute of Technology, Karlskrona, Sweden

***Abstract*— *This paper presents a single channel speech enhancement technique based on sub-band modulator Kalman filtering for laryngeal (normal) and alaryngeal (Esophageal speech) speech signals. The noisy speech signal is decomposed into sub-bands and subsequently each sub-band is demodulated into its modulator and carrier components. Kalman filter is applied to modulators of all sub-bands without altering the carriers. Performance of the proposed system has been validated by Mean Opinion Score (MOS) for laryngeal and Harmonic to Noise Ratio (HNR) for alaryngeal speech. An improvement of 20% has been observed in MOS over sub-band Kalman filtering for laryngeal speech, while 3 to 4 dB enhancement in HNR has been observed for alaryngeal speech over the full-band Kalman filtering.*

Keywords*— *Kalman filter, Autoregressive, speech enhancement

I. Introduction

Speech enhancement is an important branch of speech signal processing that aims at suppression of noise to make a speech signal more intelligible. An enhanced version of a speech signal is useful for speech recognition applications, mobile communication and coding etc. There has been many algorithms proposed for speech enhancement including but not limited to spectral subtraction [1], [2], Wiener filtering [3], adaptive gain equalizer [4], [5], [6], [7] and Kalman filtering [8], [9].

Kalman filtering is considered to be an optimal speech enhancement algorithm that relies on a Minimun Mean Square Error (MMSE) [10], [8] based method. The Kalman filtering based speech enhancement has several advantages over other speech enhancement methods, e.g. speech production model using Linear Predication (LP), inherited to Kalman filtering modeling. Kalman filter produces optimum results for non-stationary signals and do not need stationary condition like Wiener filtering [10].

The Kalman filter is used for single channel speech enhancement by Analysis-Modification-Synthesis (AMS) frame work, where noisy speech signal is segmented into frames using short time Fourier transform (STFT), then a modification of amplitude of STFT is applied using Kalman filtering followed by inverse STFT and synthesis for enhanced speech signal [11]. Paliwal introduced the Kalman filtering for speech enhancement [8]. Further modification to Kalman filtering has been observed using the EM algorithm for autoregressive (AR) estimation for Kalman filtering [12], [13], [14]. The enhancement for colored noise corrupted speech has also been investigated in [15] using Kalman filter. The most important and less complex modification done by sub-band based Kalman filtering for speech enhancement is by dividing the speech signal into a number of sub-bands followed by Kalman filtering of each sub-band [16], [17].

The Esophageal (E) speech is one type of alaryngeal speeches used for speech production after laryngeal cancer treatment, where larynx has been removed and normal speech in no more possible. The E speech has low quality due to irregular vibration of Paryngo-esophageal (PE) segments, and enhancement of E speech has been extensively treated by LPC analysis/synthesis [18], [19], [20], [21], [22], statistical methods [23], [24], [25] and detailed analysis of E speech by our group can be consulted from [26], [27], [28], [29], [30], [31], [32]. The Kalman filter has been used for enhancement of E speech along with pole stabilization and, improvement observed over LPC analysis/synthesis framework [33], [34].

Recent research has used the approach to model speech signals as the combination of low and high frequency components, called modulators and carriers respectively. The modulators (low frequency) are considered to be most important for speech intelligibility, i.e. if speech modulators are replaced by a constant value, while preserving carriers, unintelligible speech is obtained, in comparison to the case of preserving modulators and replacing carriers with constant value retains the intelligibility of speech [35]. Mathematically,

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

where $m(n)$ and $c(n)$ are modulators and carriers respectively. A trend has been observed in recent years that speech enhancement by modifying modulators of speech signal is done using different techniques. Results justify the use of modulator filtering, e.g. convex optimization, and center of gravity (CoG) demodulation, used to enhance speech signals [36], [37].

This paper introduces a modification to sub-band based Kalman filter based speech enhancement [16], by decomposing sub-bands into its modulators and carriers components. The Kalman filter is applied to modulators of sub-bands instead of sub-bands directly. Performance of the system has been validated by Mean Opinion Score (MOS) and spectrogram for laryngeal (normal) speech by comparing it to sub-band Kalman filtering [16], and Harmonic to Noise Ratio (HNR) used for alaryngeal (E speech) by comparing it with full-band Kalman filtering E speech enhancement [33], [34]. The next sections introduce system components followed by results and conclusion.

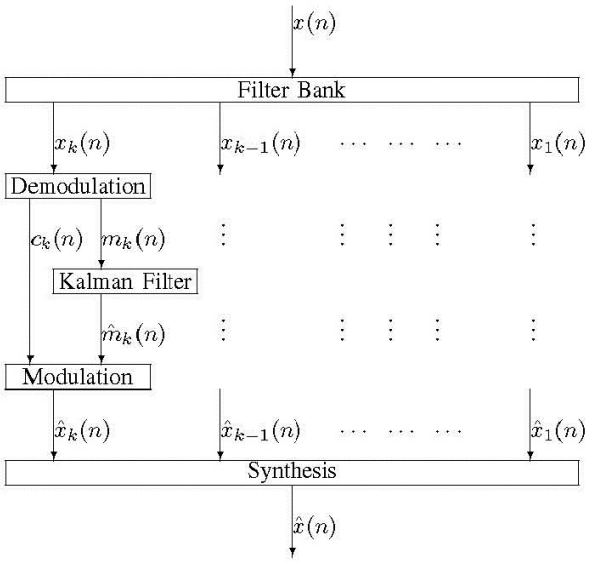

Fig. 1. Sub-band Modulator Kalman Filtering Based Speech Enhancement

II. System Design

Fig. 1 shows the proposed system used for the enhancement of noisy speech signal $x(n)$.

A K bands band-pass filter is used to divide the input speech signal $x(n)$ into sub-bands according to:

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

where $h_k(n)$ is impulse response of the kth sub-band filter and $*$ is convolution operator. Each sub-band is demodulated into modulator $m_k(n)$ and carrier $c_k(n)$ coherently according to CoG demodulation (Section III-A).

$$x_k(n) = m_k(n)c_k(n) \tag{3}$$

Sub-band modulators are modified by Kalman filtering (Section IV), given by:

$$\hat{x}_k(n) = \hat{m}_k(n)c_k(n) \tag{4}$$

where $\hat{m}_k(n)$ is modified modulator for sub-band k. The final enhanced signal is obtained by adding all the modified sub-bands according to the synthesis equation:

$$\hat{x}(n) = \sum_{k=1}^{K} \hat{x}_k(n) \tag{5}$$

III. Demodulation

Natural signals such as speech can be represented by the corresponding high frequency and low frequency components, called carriers and modulators respectively [35], [38], [39], [40]. The speech signal can be represented (in modulators and carriers sense) by equation (1). The decomposition of speech signal into $m(n)$ and $c(n)$ can be acquired coherently or non-coherently [35], [39], [40]. The non-coherent demodulation estimates the modulators and carriers independent of each other, while in coherent demodulation carriers are estimated first and then modulators are estimated based on the equation (1). In this paper, coherent demodulation has been used because of its advantages over the non-coherent and in the present case, carrier estimation is done using spectral center of gravity [41], [35], [42].

A. Spectral Center of Gravity Carrier Estimation

The demodulation framework works on sub-bands, the filter bank divides the speech signal into sub-bands, demodulation process decomposes each sub-band into its carrier and modulator components.

1) Sub-band Instantaneous Frequency: The first step in calculating the carrier is to detect the instantaneous frequency $\omega_k(n)$ of each sub-band. The center of gravity approach estimates the $\omega_k(n)$ as the average frequency of instantaneous spectrum of $x_k(n)$ [41], [35]. The instantaneous spectrum of x_k is calculated according to:

$$S_k(\omega, n) = \sum_p g(p)x_k(n+p)e^{-j\omega p} \tag{6}$$

where $g(p)$ is a window function (hamming window of length 128 is used for this experiment). Center of Gravity (CoG) estimation of $\omega_k(n)$ is given by:

$$\omega_k(n) = \frac{\int_{-\pi}^{\pi} \omega |S_k(\omega, n)|^2 d\omega}{\int_{-\pi}^{\pi} |S_k(\omega, n)|^2 d\omega} \tag{7}$$

The phase $\phi_k(n)$ is obtained by the following equation:

$$\phi_k(n) = \sum_{p=0}^{n} \omega_k(p) \tag{8}$$

2) Carrier estimation: Carrier $c_k(n)$ obtained by exponentiating $\phi_k(n)$:

$$c_k(n) = exp[j\phi_k(n)] \tag{9}$$

The carrier estimation for sub-band k gives the related modulator as:

$$m_k(n) = x_k(n)/c_k(n) = x_k(n)c_k^*(n) \tag{10}$$

IV. Subband Modulator Kalman Filtering

It is considered that modulators of speech signal can be represented by an autoregressive (AR) process, i.e. output of an all-pole system excited by white Gaussian noise and represented by a difference equation:

$$m_k(n) = \sum_{j=1}^{p} a_{k,j} m_k(n-j) + w_k(n) \tag{11}$$

where $a_{k,j}(n)$, p and $w_k(n)$ are Linear Predication Coefficients (LPC), order of AR process and input white Gaussian noise(with zero mean and variance $\sigma_{k,w}^2$) respectively for the

kth sub-band modulator $m_k(n)$. The observed noisy modulator for sub-band k is given by $s_k(n)$ as:

$$s_k(n) = m_k(n) + v_k(n) \tag{12}$$

where $v_k(n)$ is white Gaussian additive observation or measurement noise with zero mean and variance $\sigma^2_{k,v}$ for sub-band k. The equations given above can be given in the state space representation as:

$$m_k(n) = F_k m_k(n-1) + g w_k(n) \tag{13}$$

$$s_k(n) = H^T m_k(n) + v_k(n) \tag{14}$$

where $m_k(n) = [m_k(n-p+1) m_k(n-p+2) \cdots m_k(n)]$.

$$F_k = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ -a_{k,p} & -a_{k,p-1} & -ak,p-2 & \dots & -ak,1 \end{bmatrix} \tag{15}$$

$$g^T = H^T = [0, 0, \dots, 1] \tag{16}$$

The Kalman filter provides the estimate of $m_k(n)$, providing observation $s_k(1), s_k(2), ...s_k(n)$ [15] as:

$$\hat{m}_k(n) = F_k \hat{m}_k(n-1) + K_k(n)[s_k(n) - H^T F_k \hat{m}_k(n-1)] \tag{17}$$

$$K_k(n) = P_k(n|n-1) H [R_k + H^T P_k(n|n-1) H]^{-1} \tag{18}$$

$$P_k(n|n-1) = F_k P_k(n-1|n-1) F_k^T + g Q_k g^T \tag{19}$$

$$P_k(n) = [I - K_k(n) h^T] P_k(n|n-1) \tag{20}$$

where $K_k(n)$ is Kalman gain, $P_k(n|n-1)$ is a priori error covariance matrix and $P_k(n)$ is error covariance matrix, R_k and Q_k are measurement noise covariance matrix and input noise covariance matrix respectively for sub-band k. The system is initialized using the noisy modulator:

$$\hat{m}_k(0) = m_{k,0} = [s_k(1), s_k(2), \dots, s_k(p)] \tag{21}$$

$$P_k(0|0) = P_{k,0} = diag[R_k, R_k, \dots, R_k] \tag{22}$$

At time instant n estimated sample is given by following relationship:

$$\hat{m}_k(n) = H^T \hat{m}_k(n) \tag{23}$$

A. Parameter Estimation

The estimation of LPC coefficients and noise variances for sub-band modulators is necessary for optimized results of Kalman filter. These parameters of each sub-band are calculated based on EM algorithm given in [12] and it is given below breifly:

- Noisy only segment from modulator of sub-band k is detected, and additive observation noise $\sigma^2_{k,v}$ is estimated.
- LPC parameters $a_{k,n}$ and variance $\sigma^2_{k,nmodulator}$ are calculated for noisy speech modulator.
- Input noise variance is estimated by $\sigma_{k,w} = \sigma_{k,nmodulator} - \sigma^2_{k,v}$

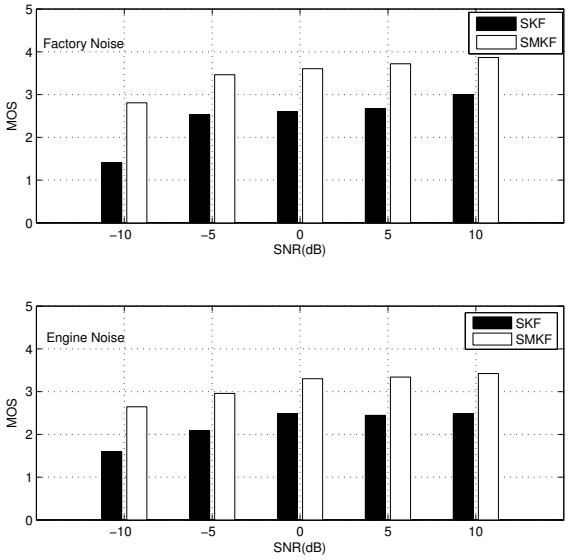

Fig. 2. Mean Opinion Score for Sub-band Kalman filter (SKF) and Sub-band Modulator Kalman Filter (SMKF).

- Kalman filter is implemented with noisy parameters, then enhanced version of modulator is used to estimate $a_{k,n}$, iterated until optimal estimate is obtained. In our work, the number of iteration are 3 as stated in [12].

V. Comparative Performance Analysis

A. Laryngeal Speech

Performance of the system has been tested using female speech signal sampled at 16 KHz, and corrupted by factory and engine noise signals with different Signal to Noise Ratio (SNR) (-10, -5, 0, 5, 10 dB). The number of filters in the filter bank effects the results, for this work, the number of filter used are 64 which gave better results in reducing residual noise. The Kalman filter uses the LPC order p of 10, and window size and step sizes are 30 and 15 millisecond respectively. This paper presents the comparison between systems based on MOS and spectrogram.

1) Mean Opinion Score (MOS): Fig. 2 shows the comparison of enhanced version speech signal with Sub-band Kalman Filtering (SKF) and Sub-band Modulator Kalman Filtering (SMKF) for MOS values. A maximum of 20% improvement can be observed and SMKF outperforms SKF for all SNR cases.

2) Spectrogram: Fig.3 and 4 show the spectrogram of speech signal corrupted by engine noise and factory noise at -10dB SNR. Although SMKF shows some loss in formants in upper frequencies but in comparison to SKF, there is less residual noise in enhanced speech signal. Significant improvement can be observed in factory noise corrupted speech signal.

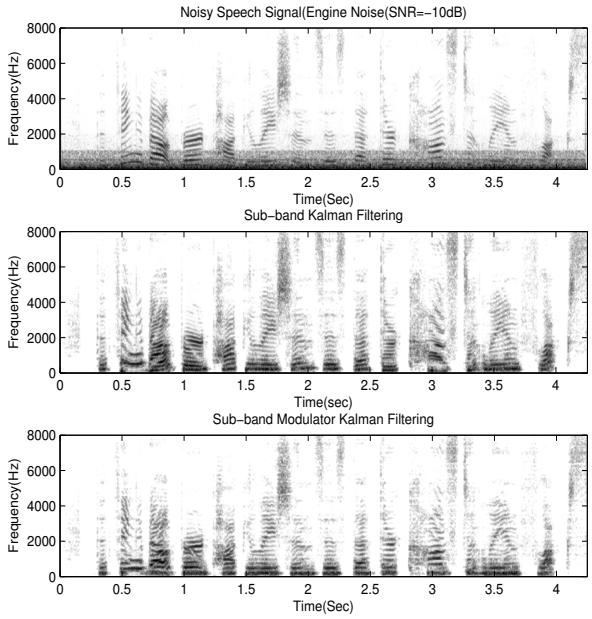

Fig. 3. Spectrogram of noisy and enhanced speech signals by systems(Engine Noise)

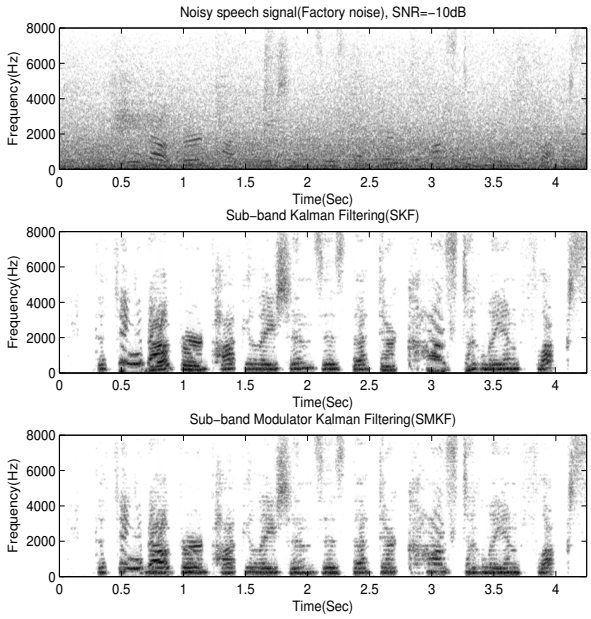

Fig. 4. Spectrogram of noisy and enhanced speech signals by systems(Factory Noise)

B. Alaryngeal Speech

The E speech vowels \a\, \e\,\i\, \o\, \u\, and \bodega\have been used for this experiment, which are recorded from alaryngeal speech rehabilitation center (6 persons) with the sampling frequency of 44100 Hz and down-sampled to 16000 Hz for computational efficiency.

1) Harmonic to Noise Ratio (HNR): VoiceSauce [43] was used to calculate HNR, with following settings, fundamental frequency range: 60 to 120 Hz (E speech fundamental frequency range is in between 60-120), frame length and overlap was set to 30 and 15 millisecond respectively and LPC order was set to 12. Fig. 5 shows the improvement of around 4 dB over the full-band Kalman filtering [33], [34], and 2 dB over sub-band Kalman filtering.

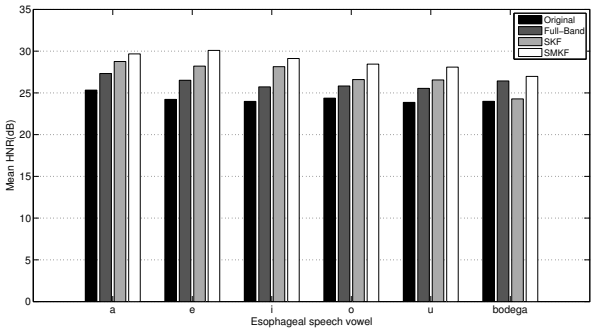

Fig. 5. Mean Harmonic to Noise Ratio (HNR)

VI. Conclusion

The modification to sub-band Kalman filtering by applying Kalman filter to modulators of sub-band by coherent decomposition has been successfully implemented for noisy laryngeal and alaryngeal speeches (E speech). Results thus obtained show improvement in speech enhancement while Kalman filtering is used in modulator domain in comparison to its traditional use. The improvement in MOS and spectrogram has shown the system capability of the proposed for reducing noise from noisy laryngeal speech, and HNR improvement has confirmed the system performance over the previous methods for E speech. The future work can be the utilization of other demodulation process, e.g. non-coherent demodulation and convex optimization demodulation [36], [44], [45].

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Accoust. Speech and Sig. Proc.*, vol. 27, no. 2, pp. 113–120, 1979.

[2] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, May 2010.

[3] M. H. Hayes, *Statistical Digital Signal Processing and Modeling*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1996.

[4] M. Dahl and B. Sallberg, "Speech enhancement implementations in the digital, analog and hybrid domain," *Swedish System on Chip Conference*, 2005.

[5] N. Westerlund, M. Dahl, I. Claesson, B. Sallberg, and H. Akesson, "Analog circuit implementation for speech enhancement purposes," *Asilomar Conference on Circuits, Systems and Computers.*, 2004.

[6] M. Dahl, I. Claesson, B. Sallberg, and H. Akesson, "A mixed analog -digital hybrid for speech enhancement purposes," *ISCAS.*, 2005.

[7] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for

personal communication using an adaptive gain equalizer," *Elsevier Signal Processing.*, vol. 85, pp. 1089–1101, 2005.

[8] K. K. Paliwal and A. Basu, "A speech enhancemet method based on kalman filtering," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, pp. 177–180, 1987.

[9] B. K. J. D. Gibson and S. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. , Signal Processing*, vol. 39, no. 8, pp. 1732–1742, 1991.

[10] S. So and K. K. Paliwal, "Supressing the influence of additive noise on the kalman gain for low residual noise speech enhancement," *Elsever, Speech communication 53*, vol. 53, pp. 355–378, 2011.

[11] ——, "Modulation-domain kalman filtering for single-channel speech enhancement," *Speech Commun.*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[12] D. Weixiu and P. Driessen, "Speech enhancement based on kalman filtering and em algorithm," *IEEE Pacific Rim Conf. on Communication, Computers and Signal Processing, 1991*, pp. 142–145, 1991.

[13] S. So and K. K. Paliwal, "A long state vector kalman filter for speech enhancement," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, 2008, pp. 391–394.

[14] E. G. M. Gabrea and M. Najim, "A single microphone kalman filter-based noise cancellor," *IEEE Signal processing Letters*, vol. 6, no. 3, pp. 55–57, 1999.

[15] D. J. G. B. Koo and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Processing*, no. 8, pp. 1732–1742, 1991.

[16] W.-R. Wu and P.-C. Chen, "Subband kalman filtering for speech enhancement," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 45, no. 8, pp. 1072–1083, 1998.

[17] C. H. You, S. N. Koh, and S. Rahardja, "Subband kalman filtering incorporating masking properties for noisy speech signal," *Speech Commun*, vol. 49, no. 7-8, pp. 558–573, Jul 2007.

[18] M. Kenji and H. Noriyo, "Enhancement of esophageal speech using formant synthesis," *Acoustics, Speech and Signal Processing, International conf.*, pp. 81–85, 1999.

[19] M. Alfredo, P. Hector, T. Jorge, and O. Patricia, "Analysis and recognition of esophageal speech," *Symposium on Signal Processing and Information Technology*, vol. 5, pp. 101–106, 2006.

[20] Y. Qi, "Replacing tracheoesophageal voicing source using lpc synthesis," *Acoustical Society of America*, vol. 5, pp. 1228–1235, 1990.

[21] R. Sirichokswad, P. Boonpramuk, N. Kasemkosin, P. Chanyagorn, W. Charoensuk, and H. H. Szu, "Improvement of esophageal speech using lpc and lf model," *Internation Conf. on Biomedical and Pharamaceutical Engineering 2006*, pp. 405–408, 2006.

[22] Q. Yingyong, W. Bernd, and B. Ning, "Enhancement of female esophageal and tracheoesophageal speech," *Acoustical Society of America*, vol. 98(5, Pt1), pp. 2461–2465, 1995.

[23] K. T. T. S. H. S. K. Doi, H.; Nakamura, "Statistical approach to enhancing esophageal speech based on gaussian mixture models," *Acoustics Speech and Signal Processing(ICASSP), 2010 IEEE International Conference*, pp. 4250–4253, 2010.

[24] M. Kenji, H. Noriyo, K. Noriko, and H. Hajime, "Enhancement of esophageal speech using formant synthesis," *Acoustic. Sci. and Tech.*, pp. 69–76, 2002.

[25] M. P.-M. H. Razo-Chavez, A.; Nakano-Miyatake, "An alaryngeal speech enhancement method based on adpcm approach," *Circuits and Systems, MWSCAS'09, IEEE, International Midwest Symposium*, pp. 1097–1101, Auguest 2009.

[26] B. Garcia, J. Vicente, A. Alonso, and E. Loyo, "Esophageal voices: glottal flow restoration," *Acoustics, Speech and Signal Processing 2005(ICASSP 05)*, pp. 141–144, 2005.

[27] A. Isasi, B. Garcia, and A. M. Zorrilla, "Corrective algorithm for esophageal voice cycle detection," *IEEE*, pp. 150–155, 2011.

[28] M. O. John and B. Garcia, "Quantifiying paramters of a source filter model for oesophageal speech," *IEEE*, pp. 532–53, 2011.

[29] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Oesophageal voice acoustic parameterization by means of optimum shimmer calculation," *WSEAS Trasactions on Systems*, pp. 489–499, 2008.

[30] B. Garcia, I. Ruiz, J. Vicente, and A. Alonso, "Formants measurement for esophageal speech using wavelet with band and resoultion adjustment," *IEEE Symposium on Signal Processing and Information Technology*, pp. 320–325, 2006.

[31] B. Garcia, J. Vicente, I. Ruiz, A. Alonso, and E. Loyo, "Improvement of esophageal voice's pitch," *Proc. of the 7th Int. Conference on Digital Audio Effects(DAFx04), Italy*, pp. 307–310, 2004.

[32] B. Garcia, I. Ruiz, A. Mendez, and M. Mendezona, "Objective characterization of oesophageal voice supporting medical diagnosis rehabilitation and monitoring," *Computers in Biology and Medicine, Elsevier*, pp. 97–105, 2009.

[33] B. Garcia and A. Mendez, "Oesophageal speech enhancement using poles stablization and kalman filtering," *ICASSP*, pp. 1597–1600, 2008.

[34] O. Ibon, B. Garcia, and Z. M. Amaia, "New approach for oesophageal speech enhancement," *10th International conference, ISSPA*, vol. 5, pp. 225–228, 2010.

[35] S. Schimmel, "Theory of modulation frequency analysis and modulation filtering with applications to hearing devices," Ph.D. dissertation, University of Washington, 2007.

[36] R. Ishaq, M. Shahid, B. Lovstrom, B. G. Zapirain, and I. Claesson, "Modulation frequency domain adaptive gain equlizer using convex optimization," *6th International Conference on Signal Processing and Communication Systems- 2012*, 2012.

[37] M. Shahid, R. Ishaq, B. Sallberg, N. Grbic, B. Lovstrom, and I. Claesson, "Modulation domain adaptive gain equalizer for speech enhancement," in *Signal and Image Processing Application 2011, by IASTED*, 2011.

[38] S. Schimmel and L. E. Atlas, "Analysis of signal reconstruction after modulation filtering," *Advanced Signal Processing Algorithms, Architectures, and Implementations*, vol. 5910, pp. 163–172, 2005.

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

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

[41] P. Clark and L. E. Atlas, "Time-frequency coherent modulation filtering of non-stationary signals," *IEEE transaction on Signal Processing*, vol. 45, no. 57, pp. 4323–4332, 2009.

[42] P. C. Les Atlas and S. Schimmel, "Modulation toolbox version 2.1 for matlab," http://isdl.ee.washington.edu/projects/modulationtoolbox/, September 2010.

[43] A. Alwan. (2012, Feb.) Voicesauce: A program for voice analysis @ONLINE. [Online]. Available: http://www.ee.ucla.edu/ spapl/voicesauce/

[44] S. Boyd and L. Vandenberghe, *Convex Optimization*, 1st ed. Cambridge, UK: Cambridge University Press, 2004.

[45] G. Sell and M. Slaney, "Solving demodulation as an optimization problem," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2051 –2066, nov. 2010.