International Journal of Computer Applications (0975 - 8887)
Volume 181 - No.33, December 2018

Wiener Filter in Wavelet Domain for Mel-LPC based Noisy Speech Recognition

M. Babul Islam
Dept. of Applied Physics and Electronic Engineering
University of Rajshahi, Rajshahi 6205, Bangladesh

Md. Hamidul Islam
Dept. of Applied Physics and Electronic Engineering
University of Rajshahi, Rajshahi 6205, Bangladesh

Md. Monsur Rahman
Faculty of Science and Engineering
North Bengal International University
Rajshahi 6100, Bangladesh

ABSTRACT

This paper deals with a wavelet domain Wiener filter to estimate enhanced Mel-LPC spectra in presence of additive noises. In this implementation, Daubechies 4 (db4) wavelet function has been used as mother wavelet which enables a fast computation and decomposition using perfect reconstruction of filterbank. To implement the filter, noise is estimated from the initial 20 frames of input speech signal without applying any voice activity detection (VAD) system. In the proposed system, filtering is done in wavelet domain using Wiener gain. After filtering, inverse wavelet transform is applied to obtain enhanced time domain speech signal. Using this enhanced speech signal Mel-LP cepstral coefficients are calculated as speech feature. The proposed system is evaluated on Aurora-2 database and it has been found that the Wiener filter improves the overall word accuracy from 58.66 to 75.88% and the average Aurora-2 relative improvement has been found to be 42.50% for test set A.

Keywords

Wiener filter, Wavelet Transform, Mel-LPC, Noisy speech recognition, Aurora-2 database

1. INTRODUCTION

Currently speech recognition systems have been using widely in real world applications. In laboratory conditions, the performance of speech recognition system is quite satisfactory. However, performance severely degrades in real environments because of noises. Environmental noises contaminate speech signal and change the data vectors representing the speech, for instance, reduce the dynamic range, or variance of cepstral coefficients within the frame. As a result, a serious mismatch is occurred between the acoustic model of training and test conditions. This may potentially degrade the performance of recognition systems. Hence, noise robustness is an important issue for automatic speech recognition.

In real environment, it is obvious that the speech and noise are not separately available rather it is a composite signal, and unbiased power estimation for both speech and noise is difficult. Since most of the filtering based techniques primarily depend on power estimation, it is very hard to recover the clean speech from the noisy signal [1]. However, many researchers tried to improve the power estimation by deploying efficient voice activity detector [2], [3], [4].

Recently, Wavelet Transform (WT) has been applied in different speech processing applications as an efficient signal processing tool taking advantages of its excellent time-frequency resolution. The WT enables better frequency resolution at low frequencies and better time localization of the transient phenomena of speech signal in the time domain. Consequently, it resembles to the first stage of human auditory perception and to basilar membrane excitation where the WT introduces almost logarithmic frequency sensitivity.

There are many techniques to enhance the noisy speech signal based on the additive property of noises. The widely used methods to remove the additive noise are spectral subtraction with many variants [5], [6] and time or frequency domain Wiener filtering [2], [7], [8]. However, denoising in wavelet domain has been found to be advantageous in some seminal works [9], [10], [11] where thresholding principle was applied on wavelet decomposed speech signal.

In this paper, a wavelet domain Wiener filter has been proposed for Mel-LPC based noisy speech recognition. The Wiener gain for each frequency band has been estimated from the wavelet decomposed speech signal and then each wavelet coefficient is multiplied with the Wiener gain to estimate the enhanced coefficients. It is a very simple technique with low computational cost. In this proposed research work Daubechies 4 (db4) wavelet [12] function is used as mother wavelet because it is compactly supported wavelet that has the highest number of vanishing moments for a given support width. This also enables a fast computation and decomposition using perfect reconstruction of filterbank.

The rest of the paper is organized as follows. Section 2 gives a brief description of wavelet transform. Section 3 deals with the design of Wiener filter. An overview of Mel-LPC analysis is presented in Section 4. System overview is introduced in section 5. Experimental setup for the proposed system is given in section 6. The perfor-

mance of the proposed system is illustrated in section 7. Finally, conclusion is presented in section 8.

2. WAVELET TRANSFORM

The wavelet transform of a continuous square integrable function $x(t)$ is defined as

$$X_\psi(\nu,\tau) = \int_{-\infty}^{\infty} x(t)\psi_{\nu,\tau}(t)dt \tag{1}$$

where $\psi_{\nu,\tau}(t)$ is a wavelet function, defined by

$$\psi_{\nu,\tau}(t) = \frac{1}{\sqrt{\nu}}\psi\left(\frac{t-\tau}{\nu}\right) \tag{2}$$

and ν, τ are the scaling and shifting parameters, respectively. The function $x(t)$ can be obtained from $X_\psi(\nu,\tau)$ using the inverse wavelet transform:

$$x(t) = \frac{1}{C_\psi}\int_0^\infty \int_{-\infty}^{\infty} X_\psi(\nu,\tau)\frac{\psi_{\nu,\tau}(t)}{\nu^2}d\tau d\nu \tag{3}$$

where

$$C_\psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega}d\omega \tag{4}$$

and $\Psi(\omega)$ is the Fourier transform of $\psi(t)$.

3. WIENER FILTER

Let $x(n)$ be the input speech signal contaminated by additive noise $w(n)$, then

$$x(n) = s(n) + w(n) \tag{5}$$

where $s(n)$ is the clean speech.

Now, we define an M-th order Wiener filter in spectral domain as follows:

$$H(z) = \sum_{n=0}^{M-1} h(n)z^{-n} \tag{6}$$

The estimated clean speech based on the filter $H(z)$ is given by

$$\hat{s}(n) = \sum_{k=0}^{M-1} h(n)x(n-k) \tag{7}$$

In the spectral domain, Equation (7) can be written as

$$\hat{S}(\omega) = H(\omega)X(\omega) \tag{8}$$

In wavelet domain, the transfer function for the Wiener filter can be defined as

$$H(\psi) = \frac{S(\psi)}{X(\psi)} = \frac{S(\psi)}{S(\psi)+W(\psi)} \tag{9}$$

where $S(\psi)$, $X(\psi)$ and $W(\psi)$ are the wavelet coefficients of clean speech, noisy speech and noise signal respectively.

In terms of power, the above equation can be rewritten as

$$H(\psi) = \frac{S^2(\psi)}{S^2(\psi)+W^2(\psi)} \tag{10}$$

Therefore, in wavelet domain, the Wiener gain at band j [13] can be expressed as

$$k_j = \frac{S_j^2(\nu,\tau)}{S_j^2(\nu,\tau)+W_j^2(\nu,\tau)} \tag{11}$$

In some input frames, the estimated noise power might be higher than the noisy speech power and Wiener gain k_j will be negative. So, in our implementation, we have estimated the gain as follows:

$$k_j = max\left(\frac{X_j^2(\nu,\tau) - \overline{W}_j^2(\nu,\tau)}{X_j^2(\nu,\tau)}, 0.1\right) \tag{12}$$

where $X_j^2(\nu,\tau)$ is the wavelet power at band j for noisy speech and $\overline{W}_j^2(\nu,\tau)$ is the average wavelet noise power estimated over first 20 frames of input speech signal.

Finally, the estimated clean speech is obtained in wavelet domain as follows:

$$\hat{S}_j(\nu,\tau) = k_j X_j(\nu,\tau) \tag{13}$$

In Equation (13), the Wiener weight k_j plays the role for the degree of suppression of the contaminant to the observed signal at band j.

4. OVERVIEW OF MEL-LPC ANALYSIS

In the Mel-LPC analysis, the following all-pole model is defined for frequency warped speech signal $\tilde{x}[n]$ $(n = 0, 1, .., \infty)$ which is bilinear transformed [14] from a windowed input speech signal $x[n]$ $(n = 0, 1, .., N-1)$:

$$\tilde{H}_\alpha(\tilde{z}) = \frac{\tilde{\sigma}_e}{1+\sum_{k=1}^{p}\tilde{a}_k\tilde{z}^{-k}} \tag{14}$$

where $\tilde{z}^{-1}$ is a first-order all-pass filter,

$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha\cdot z^{-1}} \tag{15}$$

$0 < \alpha < 1$ is treated as frequency warping factor. $\tilde{a}_k$ is the k-th mel-prediction coefficient and $\tilde{\sigma}_e^2$ is the residual energy [15], [16].

The mel-prediction coefficients $\{\tilde{a}_k\}$ can be calculated directly from the input speech signal without applying bilinear transformation as shown in [2], [16].

5. SYSTEM OVERVIEW

The block diagram of the proposed system is shown in Figure 1. The mean noise power is calculated from initial 20 frames of input speech signal. Using the estimated noise power and noisy speech power Wiener gain is calculated by Equation (12). Then filtering is done in wavelet domain to obtain the enhanced details and approximate coefficients of speech signal. Finally, inverse wavelet transform is applied to obtain time domain enhanced speech signal and mel-cepstrum is calculated to use as speech feature.

6. EXPERIMENTAL SETUP

To estimate the Wiener gain for each speech frame, single level db4 wavelet function has been used as mother wavelet. Prior to wavelet transform, each speech frame was windowed using Hamming window of length 20 ms with 10 ms frame period. Before windowing, speech signal was preemphasized with a preemphasis factor of 0.97. The recognition experiment was conducted with a 12th order

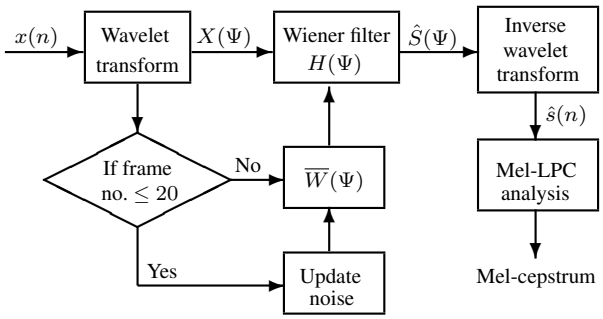

Fig. 1. Wavelet domain Wiener filtering with Mel-LPC analysis.

Mel-LPC analysis. The warping factor was set to 0.4. Each feature vector consists of 14 mel-cepstral coefficients and their delta coefficients including 0th terms.

The reference recognizer was based on HTK (Hidden Markov Model Toolkit) software package. The HMM was trained on clean condition. The digits are modeled as whole word HMMs with 16 states per word and a mixture of 3 Gaussians per state using left-to-right models. In addition, two pause models sil and sp are defined. The sil model consists of 3 states. This HMM models the pauses before and after the utterance. A mixture of 6 Gaussians models each state. The second pause model sp is used to model pauses between words. It consists of a single state, which is tied with the middle state of the sil model.

Table 1. Recognition accuracy without Wiener filter for test set A.

SNR	Noise				**Average**
	Subway	Babble	Car	Exhibition	
clean	98.74	98.55	98.51	98.80	**98.65**
20 dB	96.99	89.57	95.32	96.36	**94.56**
15 dB	92.72	72.64	82.37	92.38	**85.03**
10 dB	77.86	46.52	53.95	75.69	**63.51**
5 dB	48.88	21.46	24.04	44.80	**34.80**
0 dB	22.32	7.53	11.96	19.87	**15.42**
–5 dB	11.36	4.11	8.65	12.06	**9.05**
Average (20 to 0dB)	**67.75**	**47.54**	**53.53**	**65.82**	**58.66**

Table 2. Recognition accuracy with Wiener filter for test set A.

SNR	Noise				**Average**
	Subway	Babble	Car	Exhibition	
clean	98.77	98.46	98.57	98.80	**98.65**
20 dB	97.18	96.40	98.06	97.04	**97.17**
15 dB	94.57	93.23	96.21	92.87	**94.22**
10 dB	85.97	84.61	90.81	81.89	**85.82**
5 dB	66.84	64.87	73.90	57.82	**65.86**
0 dB	35.98	33.77	42.59	33.01	**36.34**
–5 dB	14.98	13.48	18.70	17.28	**16.11**
Average (20 to 0dB)	**76.11**	**74.58**	**80.31**	**72.53**	**75.88**

Table 3. Aurora-2 relative improvement for test set A.

SNR	Noise				**Average**
	Subway	Babble	Car	Exhibition	
clean	–10.81%	–58.76%	–50.53%	–62.16%	**–45.57%**
20 dB	13.23%	61.95%	33.56%	22.11%	**32.71%**
15 dB	35.89%	75.66%	66.90%	28.49%	**51.73%**
10 dB	42.66%	70.62%	74.80%	34.64%	**55.68%**
5 dB	37.03%	54.43%	62.31%	30.76%	**46.13%**
0 dB	17.46%	29.89%	35.70%	21.88%	**26.23%**
–5 dB	4.85%	13.38%	12.74%	11.20%	**10.54%**
Average (20 to 0dB)	**29.25%**	**58.51%**	**54.66%**	**27.57%**	**42.50%**

Table 4. Comparative recognition results among WI007, Mel-LPC (MLPC) without Wiener filter and with Wiener filter for test set A.

Front-end	Noise				**Average**
	Subway	Babble	Car	Exhibition	
WI007	69.48	49.88	60.60	65.39	**61.34**
MLPC w/o WF	67.75	47.54	53.53	65.82	**58.66**
MLPC w/ WF	76.11	74.58	80.31	72.53	**75.88**

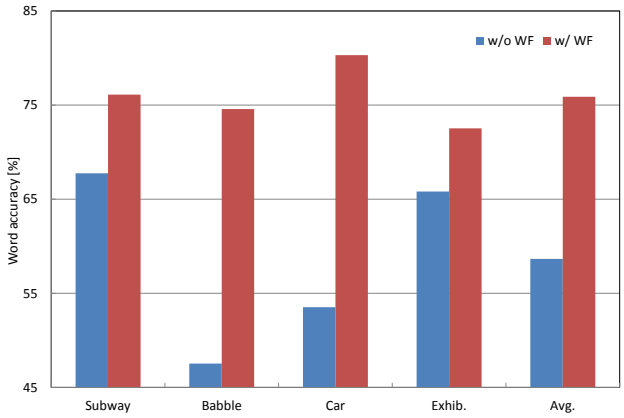

Fig. 2. A comparative result between the recognition accuracy with and without Wiener filter for noise types subway, babble, car and exhibition.

7. EXPERIMENTAL RESULTS

The performance of the proposed system was evaluated on test set A of Aurora-2 database [17], which is a subset of TIDigits database [18] contaminated by additive noises and channel effects.

The recognition accuracy with and without Wiener filter has been presented in Table 1 and Table 2. As shown in Table 1, the average recognition accuracy without applying Wiener filter is found to be 58.66% for the SNR range between 0 and 20 dB. On the contrary, with Wiener filter the overall recognition accuracy is 75.88%. It has also been observed that the most significant improvement is obtained for car noise which is 80.31% on the average. A comparison between the recognition accuracy with and without Wiener filter for each noise type is presented in Figure 2.

Finally, the Aurora-2 relative improvement is presented in Table 3. The Aurora-2 relative improvement is calculated by comparing the recognition accuracy with the baseline result obtained by applying

the Aurora WI007 front-end [19]. As shown in Table 3, the overall relative improvement obtained by the proposed Wiener filter is 42.50% for test set A. A comparative result among Aurora WI007, Mel-LPC without Wiener filter and Mel-LPC with Wiener filter is presented in Table 4.

8. CONCLUSION

In this work, a wavelet domain Wiener filter has been presented by estimating frame by frame Wiener gain to enhance speech signal contaminated by additive noises. In this implementation single level wavelet decomposition is performed using db4 wavelet function. A significant improvement has been obtained in recognition accuracy using this filter. The overall recognition accuracy has been found to be 75.88% for test set A with the proposed system and the Aurora-2 relative improvement is found to be 42.50%.

9. REFERENCES

[1] Gomez, R., et al. 2015. Optimized wavelet-domain filtering under noisy and reverberant conditions. APSIPA Transactions on Signal and Information Processing, 4.

[2] Islam, M. B., et al. 2007. Mel-Wiener filter for Mel-LPC based speech recognition. IEICE Transactions on Information and Systems, E90-D (6): 935-942.

[3] Ayat, S., et al. 2006. An improved wavelet-based speech enhancement by using speech signal features. Computers & Electrical Engineering, 32(6): 411-425.

[4] Cohen, I. 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing, 11 (5): 466-475.

[5] Boll, S. F. 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing: 27(2): 113-120.

[6] Lockwood, P. and Boudy, J. 1992. Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. Speech Communication: 11 (23): 215-228.

[7] Agarwal, A. and Cheng, Y. M. 1999. Two-Stage Mel-Warped Wiener Filter For Robust Speech Recognition. Proc. ASRU99: 67-70.

[8] Macho, D., et al. 2002. Evaluation of a noise-robust DSR front-end on Aurora databases. Proc. ICSLP: 17-20.

[9] Johnstone, I. M. and Silverman, B. W. 1997. Wavelet threshold estimators for data with correlated noise. Journal of the Royal Statistical Society: 59 (2): 319-351.

[10] Shao, Y. and Chang, C. H. 2005. A versatile speech enhancement system based on perceptual wavelet denoising. IEEE International Symposium on Circuits and Systems: 864-867.

[11] Bahoura, M. and Rouat, J. 2001. Wavelet speech enhancement based on the Teager energy operator. IEEE Signal Processing Letters: 8 (1): 10-12.

[12] Daubechies, I. 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. on Information Theory: 36(5): 961-1005, 1990.

[13] Gomez, R. and Kawahara, T. 2010. Optimizing spectral subtraction and Wiener filtering for robust speech recognition in reverberant and noisy conditions. ICASSP2010.

[14] Oppenheim, A. V. and Johnson, D. H. 1972. Discrete representation of signals. IEEE Proc., 60(6): 681-691.

[15] Strube, H. W. 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. America, 68(4): 1071-1076.

[16] Matsumoto, H., et al. 1998. An efficient Mel- LPC analysis method for speech recognition. Proc. of ICSLP98: 1051-1054.

[17] Hirsch, H. G. and Pearce, D. 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR2000: 181-188.

[18] Leonard, R. G. 1984. A database for speaker independent digit recognition. ICASSP84, 3: 42.11.1-42.11.4.

[19] ETSI standard document. 2000. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm. ETSI ES 201 108 v1.1.1 (2000-02).