LOW DISTORTION SPEECH DENOISING USING AN ADAPTIVE PARAMETRIC WIENER FILTER

Ningping Fan
fan@scr.siemens.com

Siemens Corporate Research Inc.
755 College Road East
Princeton, NJ, USA

ABSTRACT

This paper describes a parametric Wiener filter designed for noise removal with low distortion of the speech signal. The classic Wiener filter is augmented with a proportional variable for noise estimation, and a floating floor variable for the transfer function. These two variables are adaptive to the estimated noise energy in parametric relations determined experimentally for the corresponding noise estimator. The optimization of those parameters can enable the filter to achieve low distortion noise removal. Experiments using some office and home appliance noises have shown superior performance in comparison to the common Wiener filter and the spectral subtraction approaches. The proposed method has comparable quality but less computational demands than the psycho-acoustically motivated Gustafsson filter. Because of low distortions, the filter may also be used in cascading with others to achieve better total performance.

1. INTRODUCTION

Many filtering algorithms for additive noise reduction have been developed, including spectral subtraction, Wiener filtering, and psycho-acoustically motivated filters [1, 2, 3]. Recently, nonlinear solutions have been proposed on discrete wavelet transform via thresholding the DWT coefficients [4, 5]. The latter has shown that it can be advantageous to cascade a perceptually adaptive wavelet-denoising filter after a low distortion spectral subtraction filter.

Cascading multiple filters would further improve speech quality, only if each filter could keep improving. It requires that each filter produces very small distortions to avoid them multiplying, and is workable in wide SNR ranges. Cascading filters based on complementing theories can achieve a better solution by joining strength of different approaches. It may also be performed in the same domain without the transform overhead.

However, the most currently available filters do not satisfy the low distortion requirement. They remove noise but also produce distortions. They degrade speech quality outside a narrow SNR range, especially for the high SNR signals, because the less degradation due to noise interference but more due to filtering distortion. If they are used in cascading, severe speech degradation can occur.

Towards solving this problem, we designed a parametric Wiener filter, which can produces much less distortions than the standard Wiener filter, as an alternative to the preprocessing method used in [5]. It improves speech quality in a reasonably wide SNR range. The basic idea is to use adaptive variables to adjust the behavior of the Wiener filter to balance the degradation due to noise interference and due to filtering distortion.

In section 2, we formulate the noising reduction problem and give an alternative derivation of the classic Wiener filter formula. In section 3, we present the proposed parametric Wiener filter. Section 4 describes the experiments and section 5 the conclusions.

2. THE CLASSIC WIENER FILTER

For a section of input speech signal corrupted with additive noise, after pre-emphasis filtering, window function multiplication, and FFT, each data sample can still be considered as the sum of signal and noise components, because all those operations are linear.

$$X(f) = S(f) + N(f) \qquad 0 \le f \le \frac{F}{2} \qquad (1)$$

Where F is the number of FFT coefficients. Only half of them are considered due to symmetry of FFT. We want to find a filter H, so that the estimated signal

$$\hat{S}(f) = H(f)X(f) \qquad (2)$$

will minimize the following objective function.

$$J(f) = E\{(\hat{S}-S)\overline{(\hat{S}-S)}\} = E\{(HX-X+N)\overline{(HX-X+N)}\} \quad (3)$$

To minimize J, we partial derivative it with respect to H, and let it equal to zero.

$$\frac{\partial J}{\partial H} = 2E\{(HX-X+N)\frac{\partial\overline{(HX-X+N)}}{\partial H}\} = 2E\{(HX-X+N)\overline{X}\} = 2(HE\{X\overline{X}\}-E\{X\overline{X}\}+E\{N\overline{X}\}) = 0 \quad (4)$$

Then, we have H to be

$$H_{opt} = \frac{E\{X\overline{X}\}-E\{N(\overline{S}+\overline{N})\}}{E\{X\overline{X}\}} = \frac{E\{X\overline{X}\}-E\{N\overline{S}\}-E\{N\overline{N}\}}{E\{X\overline{X}\}} = \frac{E\{X\overline{X}\}-E\{N\overline{N}\}}{E\{X\overline{X}\}} = \frac{R_X-R_N}{R_X} \quad (5)$$

Where $E\{N\overline{S}\}=0$ because signal and noise components are uncorrelated.

3. THE PARAMETRIC WIENER FILTER

Figure 1 shows a diagram of actual filtering operation. The filtering block is between the analysis and synthesis modules. It consists of a filter to compute the transfer function *H(f)* and multiply it with the analysis output *X(f),* and a noise estimator producing the estimation of noise variance $\hat{R}_N(f)$.

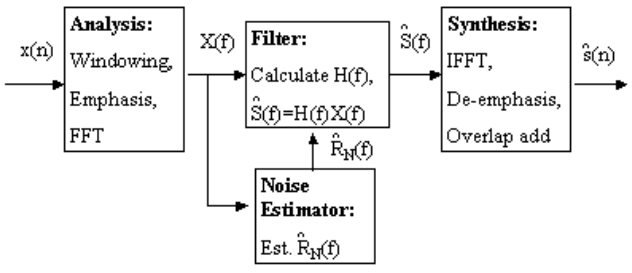

Figure 1 – The diagram of filtering operation

Let the smoothed version of input signal and noise variances be:

$$\widetilde{R}_X(f) = (1-\alpha)\widetilde{R}_X(f) + \alpha X(f)\overline{X(f)}$$
$$\widetilde{R}_N(f) = (1-\beta)\widetilde{R}_N(f) + \beta\hat{R}_N(f) \quad (6)$$

The parametric Wiener filter is given as

$$H_{pw}(f) = \max\left(\frac{\widetilde{R}_X(f)-\gamma(f)\widetilde{R}_N(f)}{\widetilde{R}_X(f)},\ h(f)\right) \quad (7)$$

where γ is the proportional variable for noise estimation, and *h* is the floating floor variable of the transfer function. The variants of (7) has been used before, where γ and *h* are constants [6]. However, our experiments have shown that for different SNR signals at different frequency bins, there are different optimal γ and *h* values.

Therefore, our contribution is to parameterize these two variables as follows:

$$\gamma(f,k) = \gamma_{\min}(k) + \frac{\gamma_0(k)}{c_0}\log((\tfrac{F}{2}+1)\widetilde{R}_N(f)+1)$$
$$h(f) = \max\left(h_{\max} - \frac{h_0}{c_0}\log((\tfrac{F}{2}+1)\widetilde{R}_N(f)+1),\ h_{\min}\right) \quad (8)$$

For a given data set and a particular noise estimator *k,* $H_{pw} = H(\gamma_{\min}(k), \gamma_0(k), c_0, h_{\max}, h_0, h_{\min})$, and those parameters can be optimized to achieve low distortion noise removal.

Intuitively the equation (8) reflects the following ideas. When noise is small, the γ will decrease and *h* will increase, so as to reduce the filtering operation and the distortion, because this is the main cause of speech degradation. When noise is strong, the reverse will happen, so as to increase filtering operation to reduce more noise because that is the main cause of degradation.

4. EXPERIMENTAL RESULTS

4.1. Implementation and Testing Data

The analysis and synthesis modules are implemented as the GSM mobile phone standard. 160 samples of pre-emphasized input data block are prefixed with last 40 samples of previous block, and multiplied by a cosine windowing function and padded 56 zeros at end for 256 sample FFT. After spectral filtering, IFFT is performed and outputs are de-emphasized. Then overlap-add is used to produce 160 samples of output data block.

Raw data were recorded in an ordinary office room with an air-conditioning fan off. 4 clean speeches and 7 noises representing room, fan, printer and open window environments, were played through a computer using one loudspeaker.

At the same time, a laptop was used to record wave files using a modified Siemens Optipoint500 phone device, which was positioned at various different locations and angles with respect to the loudspeaker. All the data were recorded in 16 KHz and 16 bits.

The clean speech recordings were then mixed with the noise recordings to mimic various additive noisy environments. The global SNR ratios of mixed signals are set to [-5dB, 0dB, 5dB, 10dB]. There are total 4*7*4 = 112 noisy mixtures.

4.2. Experimental Results

Using a small subset of the testing data, we optimized the parameters with respect to two different none-VAD based noise estimators, PSD and RM [7]. Within the SNR range of [-5dB, 10dB], the optimized values are

$$\begin{aligned} \gamma_{\min}(RM) &= 1.2, \quad \gamma_{\min}(PSD) = 4.0 \\ \gamma_0(RM) &= 3.8, \quad \gamma_0(PSD) = 16.0 \\ c_0 &= 3.3, \quad h_{\max} = 0.75 \\ h_0 &= 0.8, \quad h_{\min} = 0.2 \end{aligned} \tag{9}$$

Several noise reduction algorithms are used to process the entire testing data set, including proposed and some cascading combinations, as shown in Table 1.

Table 1 - Experimented noise reduction methods

Method	Description	Process time on Xeon 3.06GHz (ms/s)
mix	Unprocessed mixtures	0
m1	Wiener + RM	62.81
m2	Wiener + PSD	56.75
m4	Spectral Subtraction + RM	41.27
m5	Spectral Subtraction + PSD	35.19
m10	Gustafsson + RM	66.15
m11	Gustafsson + PSD	59.78
m16	Parametric Wiener + RM	37.33
m17	Parametric Wiener + PSD	30.94
m17m17	m17 cascading m17	61.88
m11m17	m11 cascading m17	90.72
m17m11	m17 cascading m11	90.72

Then the speech quality measurements of global SNR (gSNR), Itakura-Saito distance (IS), and weighted spectral slope (WSS), are calculated using processed and corresponding clean signals. Where IS and WSS measure the speech distortion, the higher value the less quality, and gSNR is vice versa [7].

Figure 2 shows a sample signal and its filtered results using some of the testing algorithms. The m2 removes most noise, but causes severe distortion indicated by high IS and WSS. The m17 reduces less noise but has least distortion among none-cascading methods. The m17m17 further enhances m17 with even less distortions. The m11 reduces slight more noise than m17 with slight more distortions. The m11m17 enhances the m11 results. The m17m11 improves m17 with slight more distortions.

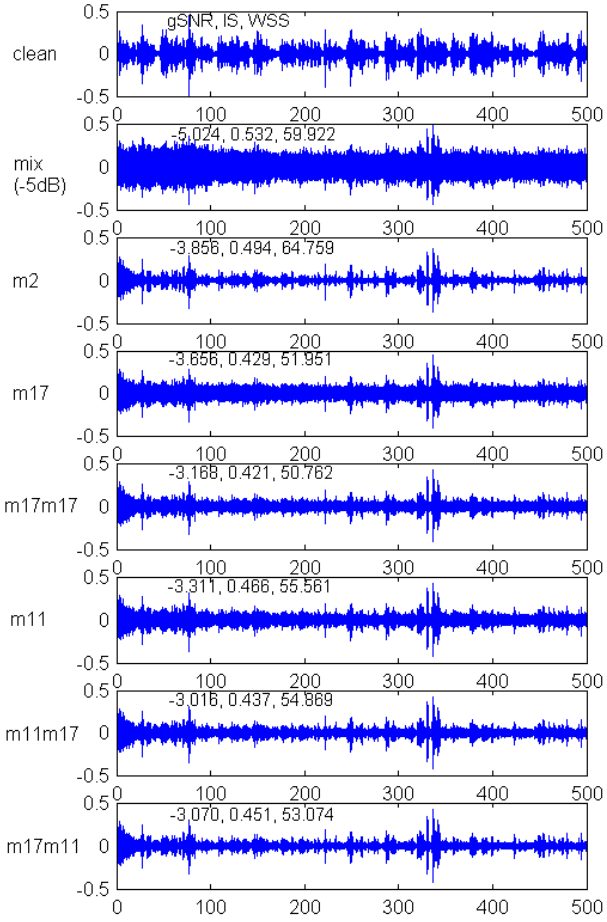

Figure 2 – A sample signal and its filtering results: The numbers above each waveform plot are the speech quality measurements for that signal.

Figure 3 shows the average scores of entire data set with respect to different SNR level. The m1, m2, m4, m5 have very high distortion measurements. The m16, m17, m17m17 are best in at higher SNR, and m11 is better in lower SNR. The m11m17 and m17m11 are better in gSNR with slightly higher distortions. Listening to the subset of processed wave files have confirmed this results.

5. CONCLUSIONS

The experiment has shown that the proposed parametric Wiener filter is among the lowest distortion filters within the tested algorithms. The standard Wiener filter and spectral subtraction approaches are the highest distortion filters within the test. The parametric Wiener filter has slightly better performance at high SNR range, while the psycho-acoustically motivated Gustafsson filter slightly better at low SNR range. Cascading filters can improve gSNR, but also increase some distortion, unless both filters have very low distortions.

Our experimental parameterization approach can also be applied to other filtering methods besides the Wiener filter. The next goal is to explore this approach with more advanced filtering technologies, such as the Gustafsson or the Wolfe filter [9], and applying cascading to achieve better total denoising quality.

The author wishes to thank J. Rosca and R. Balan at SCR for valuable discussion and suggestions.

[1] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of EUSIPCO*, 1994. Edinburgh, UK, pp. 1182-1185.

[2] W.H. Press, etl., *Numerical Recipes in C: the Art of Scientific Computing*, Cambridge Press, 1992, pp. 547-549.

[3] S. Gustafsson, P. Jax, and P. Vary, "A novel psycho-acoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proceedings of ICASSP*, 1998, pp. 397-400.

[4] C. Schremmer, T. Haenselmann, and F. Bömers, "A wavelet based audio denoiser," in *proceedings of ICME*, 2001. Tokyo, Japan, pp. 145-148.

[5] Q. Fu, and E.A. Wan, "Perceptual wavelet adaptive denoising of speech," in *Proceedings of EUROSPEECH*, 2003, Geneva, Switzerland.

[6] T.F. Quatieri, and R.B. Dunn, "Speech enhancements based on auditory spectral change," in *Proceedings of ICASSP*, 2002, pp. 257-260.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504-512, July 2001.

[8] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, NY, USA, 2000.

[9] P.J. Wolfe, and S.J. Godsill, "A perceptually balanced loss function for short-time spectral amplitude estimation," in *Proceedings of ICASSP*, 2003, pp. 425-428.

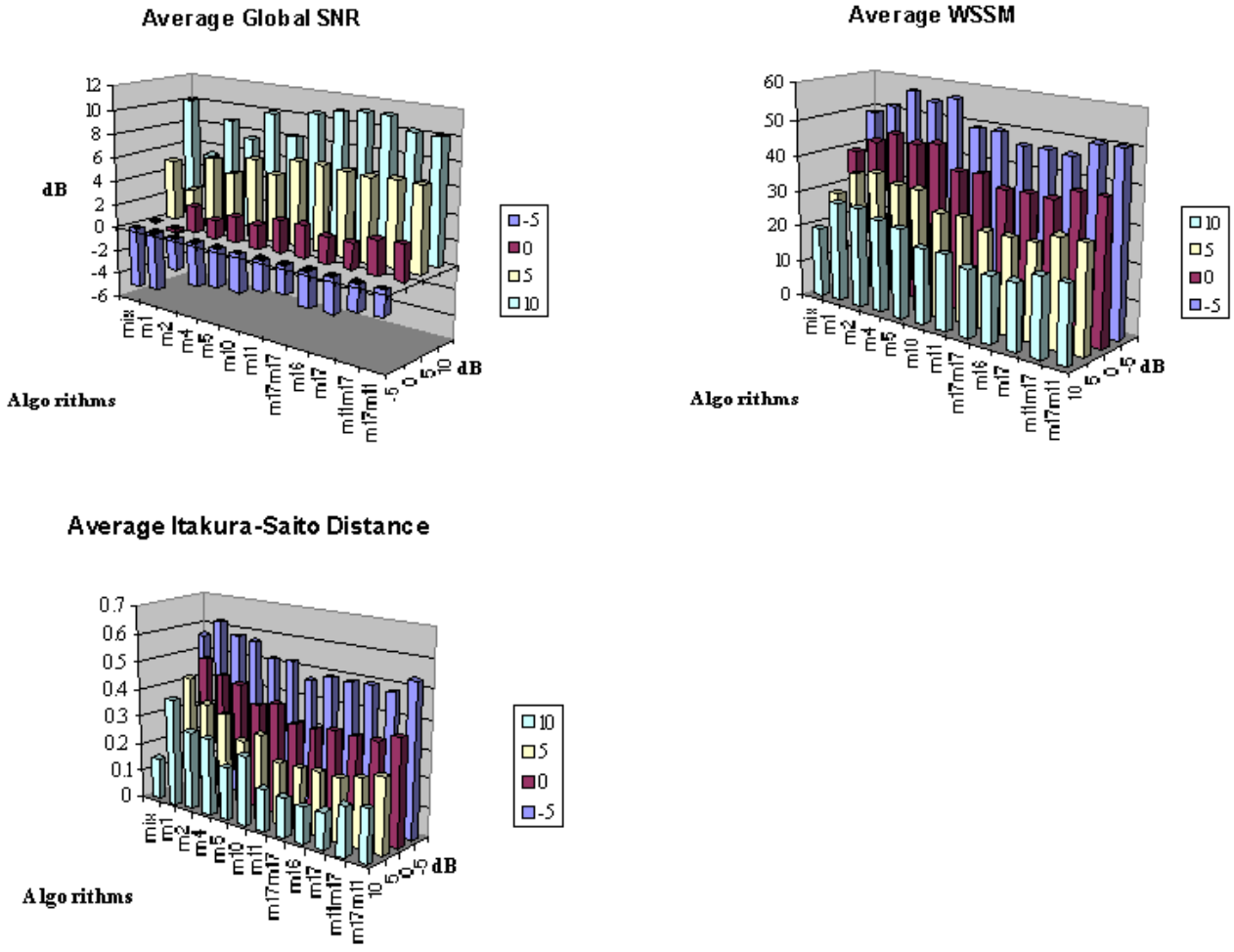

Figure 3 - Average speech quality measurements of entire testing dataset.