WAVELET TRANSFORM-BASED SPEECH ENHANCEMENT

E. Ambikairajah, G. Tattersall and A. Davis***

School of Engineering, Athlone Institute of Technology, Athlone, Ireland
*School of Information Systems, University of East Anglia, Norwich NR4 7TJ, England
**Audio and Speech Group, BT Laboratories, Martlesham Heath, Ipswich IP5 3RE, England

ABSTRACT

This paper describes a speech enhancement system using a novel combination of a Fast Wavelet Transform structure, together with "Wiener filtering" in the wavelet domain. The specific application of interest is the enhancement of speech when a cellular phone is used within a moving vehicle. Subjective tests carried out using speech with additive vehicle noise at a signal-to-noise ratio of 10 dB indicate that the Wavelet transform-based Wiener filtering approach works well. In particular, the technique was compared to several other common enhancement methods such as thresholding applied in the wavelet domain, FFT-based Wiener filtering, and spectral subtraction, and was found to outperform these other techniques.

1. INTRODUCTION

Degradation of speech quality caused by acoustic background noise is common in most speech processing applications, including mobile communications and speech recognition. Therefore, the problem of removing uncorrelated noise components from noisy speech has been widely studied in the past, and still remains an important issue in the field of speech processing research.

Classical approaches to signal enhancement include FFT-based Wiener filtering (Vaseghi, 1996) and spectral subtraction (Virag, 1995). Recently, a novel approach for denoising seismic signals using the wavelet transform was proposed by Donoho (1995). It employed thresholding in the wavelet domain and was shown to work well for signals corrupted by additive white Gaussian noise.

Subsequently, this work was extended for speech signals by Seok and Bae (1997) whose experimental results demonstrated that speech enhancement using the wavelet transform showed potential, when the speech was corrupted by additive white Gaussian noise. Their results also showed that thresholding in the wavelet domain for speech signals has the additional problem that the wavelet coefficients due to unvoiced speech could be reduced to zero after thresholding, if the thresholds are not carefully calculated. This can cause a severe reduction in the intelligibility of the reconstructed signal.

The specific application of interest in this paper is the enhancement of the speech signal when cellular phones are used within moving vehicles. Typically, the Signal to Noise Ratio (SNR) obtained from a cellular phone used within a moving vehicle is between 0 and 10 dB, resulting in increased listener effort and loss of intelligibility. Moreover, the poor signal to noise ratio in which cellular phones must operate complicates the design of the next generation of low bit rate speech coders for this application. This means that very high performance speech enhancement is required.

The most common approach to speech enhancement in non-cellular applications is based upon Wiener filtering (Vaseghi, 1996) in which short term estimates of the noise and speech signals are used to define an adaptive filter through which the noisy speech is passed, to remove as much noise energy as possible whilst simultaneously removing as little speech energy as possible. The Wiener filter approach can result in satisfactory speech enhancement but tends to distort the speech signal in perceptually unacceptable ways when the SNR is very low, as it will be in cellular applications. For example, the noise, although reduced in magnitude, is given a musical nature which is perceptually more objectionable than the original high level noise. Therefore, straightforward application of the Wiener filter approach is unsuitable for use in cellular telephony.

In this paper, a novel technique for speech enhancement in the cellular environment is described. The method uses a combination of the Wavelet transform with "Wiener filtering" in the wavelet domain. The technique was compared to several other enhancement methods such as Donoho's wavelet-domain thresholding technique and FFT-based Wiener filtering and spectral subtraction, by means of subjective tests using speech with additive vehicle noise at a signal-to-noise ratio of 10 dB. Results indicate that the proposed method provides better speech enhancement than the other techniques.

2. PROPOSED METHOD

2.1 FWT-based Speech Enhancement System

In this Section, the speech enhancement method proposed in the paper will be described in more detail. The system architecture is shown in Fig. 1. The first stage is the processing of the noisy speech signal using a Fast Wavelet Transform (FWT). A four level FWT decomposition (Strang and Nguyen, 1996) is performed resulting in five subbands covering the frequency range from 0 - 4000 Hz. The output of the fast wavelet structure is a set of "wavelet coefficients". A noisy speech input signal results in noisy wavelet coefficients at the output of the FWT decomposition. According to Donoho (1995), thresholding the wavelet coefficients has potential for recovering the original signal. However, it has been found that it is difficult to accurately estimate threshold values for all six frequency bands such that the re-synthesised speech is of acceptable quality. Various methods of calculating thresholds were investigated, but none of these proved suitable for high quality speech enhancement. Also, based on experimental results from studies of spectral subtraction, inaccurate noise threshold values result in musical noise in the reconstructed speech.

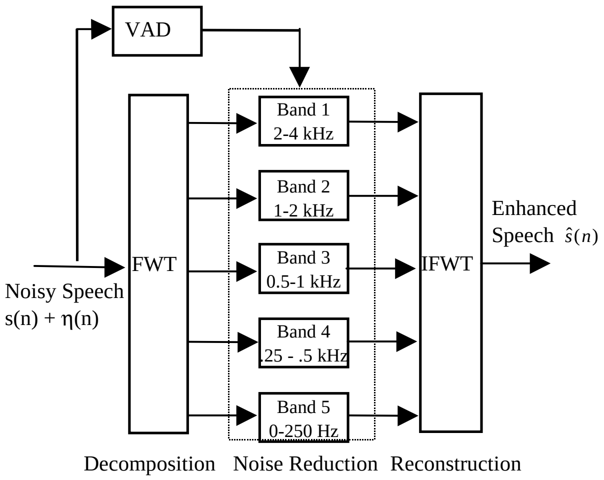

Figure 1: Speech enhancement system architecture (VAD = Voice Activity Detector)

2.2 Wavelet Denoising using Wiener Filtering

In order to avoid thresholding the wavelet coefficients, another method was investigated, whereby the wavelet coefficients were *compressed*. This method is based on "Weiner filtering" in the wavelet domain, as opposed to the well-known Weiner filtering in the frequency domain (Zelniker and Taylor, 1994). Figure 2 shows a block diagram of the conventional Wiener filtering concept.

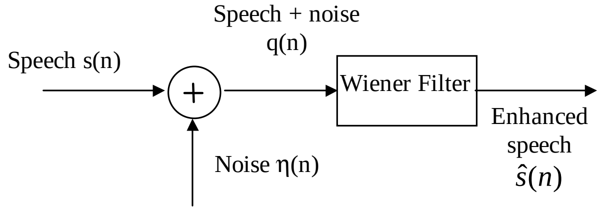

Figure 2: Block diagram of Wiener filtering paradigm

From Figure 2, we can write:

$$q(n) = s(n) + \eta(n) \tag{1}$$

and

$$\hat{s}(n) = q(n) * h(n)\,. \tag{2}$$

We can define an error signal, e(n), as follows:

$$e(n) = s(n) - \hat{s}(n) \tag{3}$$

where s(n) is the speech signal, h(n) is the Wiener filter impulse response and η(n) is the additive noise component.

By definition, a Wiener filter is a filter that minimises the mean square error, *E* (calculated as the sum of the square of the error samples over a frame). The Wiener filter coefficients h(n) are found by solving an equation in which the derivative of the mean square error with respect to the filter coefficients is set to zero as shown in Equation (4).

$$\begin{bmatrix} \frac{dE}{dh_0} \\ . \\ . \\ \frac{dE}{dh_{N-1}} \end{bmatrix} = \begin{bmatrix} 0 \\ . \\ . \\ 0 \end{bmatrix} = \begin{bmatrix} R_{ss}(0) \\ . \\ . \\ R_{ss}(N-1) \end{bmatrix} - \begin{bmatrix} R_{qq}(0,0)\ldots\ldots\ldots\ldots R_{qq}(0,N-1) \\ . \\ . \\ R_{qq}(N-1,0)\ldots\ldots. R_{qq}(N-1,N-1) \end{bmatrix} \begin{bmatrix} h_0 \\ . \\ . \\ h_{N-1} \end{bmatrix} \tag{4}$$

Equation (4) shows that the derivatives of the mean square error can be expressed in terms of the autocorrelation functions of the speech signal alone, $R_{ss}(\tau)$, and the combined speech and noise signal, R_{qq}(i,j). The equation can be re-expressed in the frequency domain using the Wiener-Khinchine relationship between power spectrum and autocorrelation function:

$$S^2(\omega) = FT\{R_{ss}(\tau)\} \tag{5}$$

Transforming Equation (4) using this result yields the well known Wiener filter frequency response.

$$H(\omega) = \frac{S^2(\omega)}{S^2(\omega) + N^2(\omega)} \tag{6}$$

where H(ω) is the transfer function of the Wiener filter, S^2(ω) is the speech power spectrum and N^2(ω) is the noise power spectrum.

Intuitively, it would appear that a "Wiener filter" could be defined in a similar way using wavelets drawn from an orthogonal set, instead of using orthogonal sinusoids. However, this only yields a Wiener filter in the true minimum mean square error sense if the Wiener Khinchine relationship applies to wavelets in the same way as to complex sinusoids. Analysis shows that the relationship applies to complex sinusoids and all similar basis functions, $u_{k,i}$ with the following property:

$$u_{k,i}.u^*_{k,p+i} = u^*_{k,p} \tag{7}$$

Where *k* is the first basis function parameter, such as frequency, *i* is the second parameter, such as time, and *p* is a shift in value of the second parameter.

Unfortunately, the Daubechies wavelets used in our experiments do not appear to have this property and so a true Wiener filter will not be obtained. However, of more practical significance is that the basis function set on which the noise removal filter is based, should have the property of discriminating between the signal subspaces occupied by the speech and the noise. The FWT appears to have this property because each wavelet is associated with an octave frequency band which matches the spectral distribution of speech energy.

Given this argument we propose simply modifying Equation (4) to operate in terms of wavelets of scale value *a*:

$$H(a) = \frac{S^2(a)}{S^2(a) + N^2(a)} \qquad (8)$$

where $S^2(a)$ and $N^2(a)$ are the speech and noise powers, calculating from the wavelet coefficients at scale *a*. The "Wiener gain" for each scale (which corresponds to a particular octave band in Figure 1) is used to modify the wavelet coefficients of the noisy signal before reconstructing the signal in the time domain by inverse wavelet transformation. For the FWT, the band is defined by an index, *i*, and so the Wiener gain, k_i, is calculated using the following equation:

$$k_i = \frac{S_i^2}{S_i^2 + N_i^2} \qquad (9)$$

where S_i^2 is the speech energy and N_i^2 is the noise energy in band *i*.

Noise segments were detected using a voice activity detector, and the noise power was calculated as follows:

$$N_i^2 = \frac{1}{L}\sum_{j=1}^{L}\left(\frac{1}{M}\sum_{l=1}^{M}\left[n_i(l,j)\right]^2\right) \qquad (10)$$

where M is the number of wavelet coefficients in a frequency band, L is the number of noisy frames over which the noise power is averaged and $n_i(l,j)$ are the wavelet coefficients associated with sub-band i for the noise input during frame j.

The signal plus noise power was calculated during speech regions as the sum of the squares of the wavelet coefficients. The noise power is then subtracted from the speech plus noise power to obtain the speech power. With these power values, the Wiener gain can be calculated for each frequency band using Equation (9). If the estimate of the noise power is greater than the estimate of the signal plus noise power, then k_i for that band may be set to zero or a small value.

If d_{ij} is the j^th noisy wavelet coefficient in band i, then the denoised wavelet coefficient is given by

$$d_{ij}(\text{denoised}) = d_{ij} \,.\, k_i \qquad (11)$$

These denoised wavelet coefficients are used to reconstruct the speech signal. When the speech power is much greater than the noise power, which would normally be the case during a voiced speech frame, then $k_i \approx 1$. If the speech and noise powers are comparable in value, then $k_i \approx 0.5$; this would normally be the case during onset and offset of voicing, or in unvoiced regions.

2.3 Wavelet Denoising using Donoho Thresholding

For comparative purposes, wavelet denoising using Donoho's thresholding technique (Donoho, 1995) was implemented. The threshold *T* was calculated using the following formula:

$$T = \sigma\sqrt{2\log(N)} \qquad (12)$$

where σ is the estimate of the standard deviation of the noise and N is the window length of the signal. For a fixed N, the threshold value is constant for all sub-bands in the wavelet domain. The noise standard deviation σ was estimated during the non-speech frames. Soft thresholding was carried out on the wavelet coefficients before reconstructing the signal.

3. RESULTS

The proposed speech enhancement algorithm has been tested as follows. In the first case, clean speech was passed through the enhancement algorithm, and, as expected, the output of the algorithm was identical to the input, i.e. the technique is transparent to speech which already has a high SNR. Clean speech samples and vehicle noise samples were added to obtain a noisy speech signal, thus simulating the use of a cellular phone in a moving vehicle. The noisy speech had a signal-to-noise ratio of 10 dB.

Non-overlapping frames of 160 samples each (20 ms) were used for analysis. The fast wavelet decomposition was obtained from each frame, using Daubechies 'db8' wavelet (Daubechies, 1990). A voice activity detector was used to determine whether the current frame was speech or non-speech. If the frame was non-speech, the estimate of the noise power in each band was updated. The Wiener gain for each band was calculated and the wavelet coefficients were scaled by the corresponding gain value for each band using Equation (11). Figure 3 shows a plot of the spectrum of one frame of voiced speech, along with the spectrum of one frame of vehicle noise. It can be clearly seen that the vehicle noise has significant low frequency content, and does not have a flat spectrum like Gaussian noise.

To illustrate the operation of the speech enhancement algorithm, Figure 4(a) shows the original speech waveform of a female speaker, sampled at 8 kHz (voiced speech segments as determined by VAD). The speech with added car noise is shown in Figure 4(b). Figure 4(c) shows the Wiener gain as a function of frame number for the lowest frequency subband, while Figure 4(d) shows the Wiener gain for the highest frequency subband. It can be seen that the Wiener gain is reduced from its maximum value of 1 during segments where the speech energy is lower.

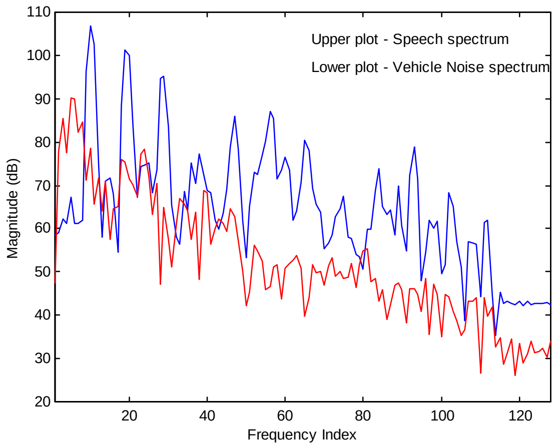

Figure 3: Spectra of voiced speech and vehicle noise.

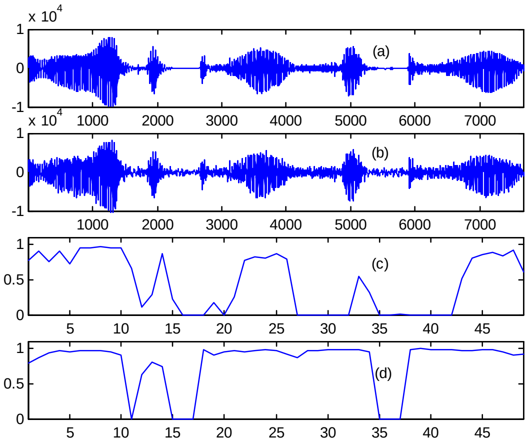

Figure 4: (a) Original speech waveform (Female speaker)
(b) Speech + Car Noise waveform
(c) Wiener gain in 0-500 Hz subband
(d) Wiener gain in 2-4 kHz subband

For comparative purposes, speech enhancement using Donoho's threshoding method, using a fixed threshold across all sub-bands, was not satisfactory. Therefore, Equation (12) was modified so that an individual threshold was calculated for each band. This modified threshold improved the quality of the enhanced speech, but noise removal not as good as Wiener filtering in the Wavelet domain.

Speech enhancement using a similar Wiener filtering technique to that used on wavelet coefficients, was performed on coefficients obtained from an FFT. In this case, the FFT spectrum was divided into five octave frequency bands, similarly spaced as for the five-band Wavelet Transform. To prevent edge effects with the FFT, speech frames were overlapped by 50% and windowed using a Hamming window. The Wiener gain per band was calculated as described above, and the FFT coefficients per band were scaled by the corresponding Wiener gain value. Again, the quality of the enhanced speech was not as good as that obtained using the proposed algorithm.

4. CONCLUSIONS

A novel speech enhancement method using Wiener filtering in the Wavelet domain has been proposed in this paper. The method has been evaluated by subjective listening tests. From these tests, the proposed approach to speech enhancement has been found to give better performance to other enhancement methods, including wavelet thresholding and FFT-based Wiener filtering. Work is in progress to implement the enhancement procedure using the Wavelet Packet Transform and auditory masking.

5. ACKNOWLEDGEMENTS

This work been carried out with funding provided by Irish Science and Technology Agency (Forbairt), and the British Council under the Collaborative Research Programme 1997.

6. REFERENCES

1. Daubechies, I. "The wavelet transform, time-frequency localization and signal analysis", IEEE Trans. Info. Theory Vol. 36, No. 3, pp. 961-1005, 1990.

2. Donoho, D. L. "De-Noising by Soft-Thresholding", IEEE Transactions on Information Theory , Vol. 41, No. 3, pp. 613-627, 1995

3. Seok, J. W. and Bae K. S. "Speech Enhancement with reduction of noise components in the wavelet domain" Proc. ICASSP'97, pp 1323-1326, 1997.

4. Strang, G. and Nguyen, T. "Wavelets and Filter Banks", Wellesley-Cambridge Press., 1996.

5. Vaseghi S. "Advanced Signal Processing and Digital Noise Reduction" , Wiley and Teubner, 1996.

6. Virag, N. "Speech Enhancement based on Masking Properties of the Auditory System", Proc. ICASSP, pp. 796-799, 1995.

7. Zelniker, G. and Taylor. F. "Advanced Digital Signal Processing", Marcel Dekker, Inc., 1994.