SUBBAND NOISE ESTIMATION FOR SPEECH ENHANCEMENT USING A PERCEPTUAL WIENER FILTER

L. Lin, W.H. Holmes and E. Ambikairajah

School of Electrical Engineering and Telecommunications
The University of New South Wales, UNSW Sydney, NSW 2052, Australia
ll.lin@ee.unsw.edu.au, H.Holmes@unsw.edu.au, Ambi@ee.unsw.edu.au

ABSTRACT

This paper proposes a fast noise estimation algorithm for speech enhancement using a perceptual Wiener filter. The noisy speech is decomposed using a critical-band-rate filterbank so that a perceptual modification of Wiener filtering can be applied in speech denoising. The subband noise estimate is updated by adaptively smoothing the noisy signal power. The smoothing parameter is chosen as a function of the estimated signal-to-noise ratio. This noise estimation technique gives accurate results even at very low signal-to-noise ratios, and works continuously, even in the presence of speech. It is effective for both non-stationary and coloured noise. Enhanced speech of good quality is obtained by the perceptual Wiener filter.

1. INTRODUCTION

Traditional speech denoising techniques are predominantly based on either Wiener filtering or spectral subtraction [5, 10]. Although these methods improve the signal-to-noise ratio, they distort the signal and also tend to introduce a perceptually annoying residual noise, often referred to as musical noise. Recent noise reduction techniques [4, 11] have exploited the masking properties of the human auditory system and have resulted in good quality speech with reduced levels of musical noise.

In most situations we have only the noisy speech signal available. The noise may be non-stationary and coloured, and its power is unknown. Its properties must be extracted from the noisy speech signal alone. Noise power estimation is crucial to effective speech enhancement. Inaccurate noise power used in the suppression rule can result in musical noise and speech distortion. Work has been carried out in recent years to find appropriate noise estimation techniques. Doblinger [3] proposed a simple algorithm for tracking the minima of the noise power. However, his algorithm cannot distinguish between a rise in noise power and a rise in speech power, so that during voiced speech intervals the noise estimates are higher than the true noise power. The minimum spectral tracking algorithm proposed by Martin [9] requires a long segment of speech to work effectively and has a large latency.

In this paper, we propose a fast and reliable noise estimation technique. The noise estimate is updated adaptively and continuously, with a smoothing parameter that depends on the estimated subband SNR. This noise estimation technique is then applied to speech denoising using the perceptually modified Wiener filter first proposed in [7, 8]. Enhanced speech of good perceptual quality is obtained.

The paper is organised as follows: Section 2 introduces the perceptually modified Wiener filter. Section 3 presents the new noise estimation technique, and section 4 gives some experimental results.

2. PERCEPTUALLY MODIFIED WIENER FILTERING FOR SPEECH DENOISING

The noisy speech $x(n)$, consisting of the clean speech $s(n)$ with the additive noise $w(n)$, is first decomposed into M bandpass signals $x_i(n)$ using a filterbank. In the implementation reported here we use the auditory filters proposed in Lin [6] to split the signal into critical band signals, although other bandpass filterbanks could be used (including the common Fourier decomposition). The auditory filters we use are approximately distortionless; that is, they satisfy Eq. 1, where $h_i(n)$ is the impulse response of the critical band filters, C is a constant and L is the number of FIR filter coefficients:

$$\sum_{i=1}^{M} h_i(n) \approx C\,\delta(L-n). \qquad (1)$$

Each noisy sub-band signal $x_i(n)$ is multiplied by the denoising gain K_i to obtain the denoised sub-band signal $\hat{s}_i(n)$. These signals are summed to yield the denoised speech $\hat{s}(n)$.

The output of the ith critical band filter (i.e. the ith noisy sub-band signal) is given by

$$x_i(n) = h_i(n) * x(n) \equiv s_i(n) + w_i(n), \qquad (2)$$

where $s_i(n) = h_i(n)*s(n)$ is the output from the ith critical band filter when the input to the filterbank is clean speech only, and $w_i(n) = h_i(n)*w(n)$ is the corresponding output when the input is noise only.

We define the signal powers $\sigma_{x_i}^2 = \mathrm{E}\{x_i^2(n)\}$, $\sigma_{s_i}^2 = \mathrm{E}\{s_i^2(n)\}$ and $\sigma_{w_i}^2 = \mathrm{E}\{w_i^2(n)\}$. The optimum channel gain K_i based on a perceptual criterion (the audible noise), is found in Lin *et al.* [7, 8] to be

$$K_i = \frac{\sigma_{s_i}^2}{\sigma_{s_i}^2 + \mu \max\{(\sigma_{w_i}^2 - \eta T_i), 0\}}, \qquad (3)$$

where T_i is the estimated masking threshold calculated using the MPEG simultaneous masking model of [1, 2], and μ and η are arbitrary parameters which add degrees of freedom to the solution. They allow a flexible trade-off between signal distortion and audible noise. We commonly choose μ to be about 1 and η to be less than 1. The classical Wiener filter is obtained

if we set $m = 1$ and $h = 0$. The generalized Wiener filter has arbitrary m, but $h = 0$.

When the noise $s^2_{w_i}$ is under the (modified) masking threshold $h\,T_i$, the gain K_i will always be 1, so that the signal is not distorted. The gain decreases as the noise exceeds this level, but it will always be larger than the optimum solutions to the Wiener and generalized Wiener problems [5], for both of which $h = 0$. It is easy to show that the speech distortion is always smaller than achieved with the Wiener solutions (i.e. if masking is not allowed for). Similarly, the noise residual is always larger than with the Wiener solution, but the difference will be less audible due to masking [7, 8].

To calculate the denoising gain in Eq. 3, both the noise power $s^2_{w_i}$ and the signal power $s^2_{s_i}$ are required. Usually only the noisy speech is available, so that the noise power and the clean signal power will have to be estimated. Here we will introduce a new noise estimation technique that gives reliable results even with strong noise.

3. ADAPTIVE NOISE ESTIMATION

We assume that noise and speech are independent non-stationary signals, but that the noise power changes relatively slowly. The subband noisy signal power, $s^2_{x_i} = E\{x_i^2(n)\}$, is estimated on a frame-by-frame basis using

$$\hat{s}^2_{x_i}(p) = \frac{1}{N}\sum_{n=0}^{N-1} x_i^2(pN+n), \qquad (4)$$

where $\hat{s}^2_x(p)$ is the estimated noise signal power calculated using frame p, and N is the frame size.

The subband noise power, $s^2_{w_i} = E\{w_i^2(n)\}$, is estimated using the one-pole smoothing filter

$$\hat{s}^2_{w_i}(p) = a_i(p)\hat{s}^2_{w_i}(p-1) + (1-a_i(p))\hat{s}^2_{x_i}(p), \qquad (5)$$

where $\hat{s}^2_{w_i}(p)$ is the estimate of subband noise power in frame p.

The smoothing parameter $a_i(p)$ at frame p is chosen as

$$a_i(p) = 1 - \min\left\{1,\ \left(\frac{\hat{s}^2_{x_i}(p)}{\bar{s}^2_{w_i}(p-1)}\right)^{-Q}\right\}, \qquad (6)$$

where Q is an integer and $\bar{s}^2_{w_i}(p-1)$ is the average or median of the noise estimates of the previous 5 to 10 frames, e.g. $\bar{s}^2_{w_i}(p-1) = 1/10\sum_{k=1}^{10}\hat{s}^2_{w_i}(p-k)$. The ratio $\hat{s}^2_{x_i}(p)/\bar{s}^2_{w_i}(p-1)$ can be considered to be an approximation to the *a posteriori* signal-to-noise ratio $s^2_{x_i}/s^2_{w_i} = (s^2_{s_i} + s^2_{w_i})/s^2_{w_i}$. Ideally a_i should be an increasing function of the ratio $s^2_{x_i}/s^2_{w_i}$. Because $\hat{s}^2_{x_i}(p)$ and $\bar{s}^2_{w_i}(p-1)$ themselves are random variables, the ratio $\hat{s}^2_{x_i}(p)/\bar{s}^2_{w_i}(p-1)$ may be smaller than 1 occasionally. The operation $\min\{\cdot\}$ is used in Eq. 6 to avoid negative $a_i(p)$ values and unstable updates.

The foregoing algorithm can be explained as follows. Firstly, if speech is absent in frame p, the new noise power calculation $\hat{s}^2_{x_i}(p)$ should be very close to the average noise estimate $\bar{s}^2_{w_i}(p-1)$, so that, from Eq. 6, $a_i(p) \approx 0$. Also, from Eq. 5 we have $\hat{s}^2_{w_i}(p) \approx \hat{s}^2_{x_i}(p)$, because of the small value of $a_i(p)$. That is, the estimate of noise power in frame p rapidly follows the power of the noisy signal in the absence of speech – there is minimal smoothing.

On the other hand, if speech is present, the new signal power $\hat{s}^2_{x_i}(p)$ is much larger than the previous noise estimate $\bar{s}^2_{w_i}(p-1)$, i.e. $\hat{s}^2_{x_i}(p) >> \bar{s}^2_{w_i}(p-1)$, so that, from Eq. 6, $a_i(p) \approx 1$. Hence the noise update in Eq. 5 is slower because of the large value of $a_i(p)$. For example, if $\hat{s}^2_{x_i}(p) = 2\bar{s}^2_{w_i}(p-1)$, which means that at this instant the speech power is equal to the noise power, we will have $a_i(p) \approx 0.94$ if $Q = 4$. The value of $a_i(p)$ increases rapidly with increasing $\hat{s}^2_{x_i}(p)/\bar{s}^2_{w_i}(p-1)$. During voiced frames we have $\hat{s}^2_{x_i}(p) >> \bar{s}^2_{w_i}(p-1)$, $a_i(p) \approx 1$, and $\hat{s}^2_{w_i}(p) \approx \hat{s}^2_{w_i}(p-1)$. That is, the noise update process almost stops and the noise estimate approximately equals that of the previous frame because the value of $a_i(p)$ is almost 1.

The integer Q controls the way in which $a_i(p)$ changes with $\hat{s}^2_{x_i}(p)/\bar{s}^2_{w_i}(p-1)$. A plot of a_i as a function of the *a posteriori* signal-to-noise ratio $s^2_{x_i}/s^2_{w_i}$ at different values of Q is shown in Figure 1. Generally, larger values of Q lead to larger values of a_i and slower noise updates, whereas smaller values of Q give faster noise updates, at the risk of possible over-estimation during long voiced intervals. The value of Q is usually chosen in the range 4 to 6.

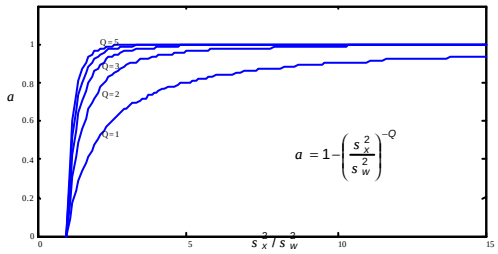


Figure 1. Plot of a as a function of the SNR s^2_x / s^2_w.

To further smooth the noise estimate, the following filtering operation is used to obtain the final noise estimate $\hat{s}^2_{w_i,final}(p)$, where a_{final} is chosen to be around 0.5-0.8:

$$\hat{s}^2_{w_i,final}(p) = a_{final}\hat{s}^2_{w_i,final}(p-1) + (1-a_{final})\hat{s}^2_{w_i}(p). \qquad (7)$$

The estimate for the subband clean signal power, $s^2_{s_i} = E\{s_i^2(n)\}$, is then calculated using

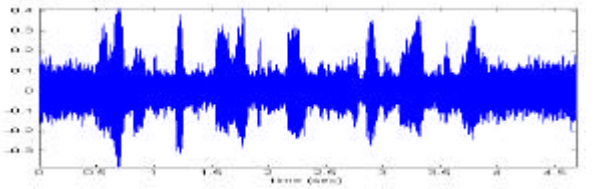
$$\hat{s}^2_{s_i}(p) = \max\left\{\hat{s}^2_{x_i}(p) - \hat{s}^2_{w_i,final}(p),\ 0\right\}, \qquad (8)$$

where $\hat{s}^2_{s_i}(p)$ is the estimate of subband clean speech power at frame p. The operation $\max\{\cdot\}$ is used to avoid negative power estimates.

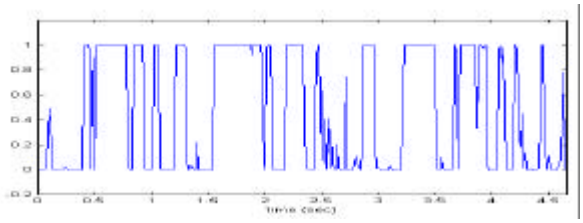
With the estimates $\hat{s}^2_{w_i,final}(p)$ and $\hat{s}^2_{s_i}(p)$ available for each subband, the masking threshold T_i and the denosing gain K_i in Eq. 3 can then be calculated on a frame-by-frame basis.

4. EXPERIMENTAL RESULTS

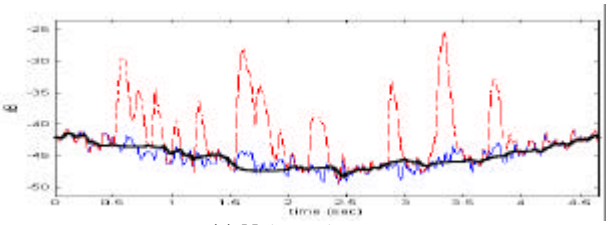
A noisy speech sentence is shown in Fig. 2a, with additive artificial white noise of slowly changing noise power (first reducing, then increasing). The overall signal-to-noise ratio for the noisy sentence is approximately 2dB. The subband noise estimate for critical band 10 (centre frequency 1170 Hz) is shown in Fig. 2c, with the dashed line being the noisy signal power; the solid thin line the true noise power, and the solid thick line the estimated noise power. The true subband noise power is calculated frame by frame using $s^2_{w_{10}}(p) = 1/N \sum_{n=0}^{N-1} w^2_{10}(pN+n)$ from the noise alone (which is known in this experiment). The value of Q in Eq. 6 is chosen to be 5, and a_{final} in Eq. 7 is 0.7.



(a) Noisy Sentence



(b) Update of the smoothing parameter $a_{10}(p)$



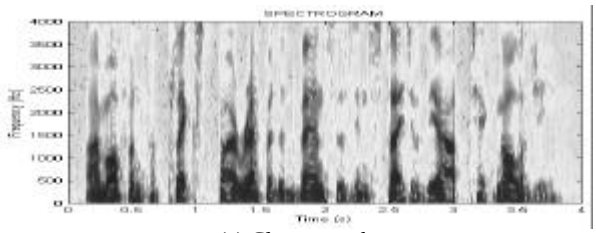
(c) Noise estimate.

Figure 2. Noise estimation in critical band 10.

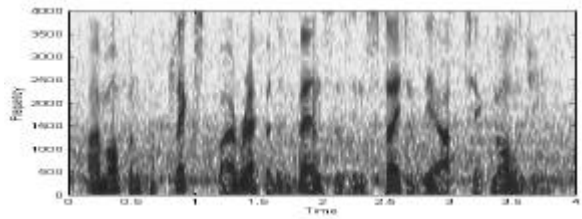
The corresponding updates of $a_{10}(p)$ are shown in Fig. 2b. As can be seen, the noise estimate follows the ideal noise power with little delay. In voiced frames, where the speech energy is high, the noise estimates almost stop changing because the value of $a_{10}(p)$ is almost 1. This behaviour can be seen in Fig. 2c in the voiced speech segment between 1.5 s and 2 s. Because of the large value of $Q = 5$, over-estimation of noise power is avoided.

Extensive testing of our noise estimation algorithm has also been made using a variety of other noises, including pink noise, tank noise, F16 noise and car noise. The proposed algorithm gives accurate and robust results at signal-to-noise ratios from -5 dB to 15 dB.

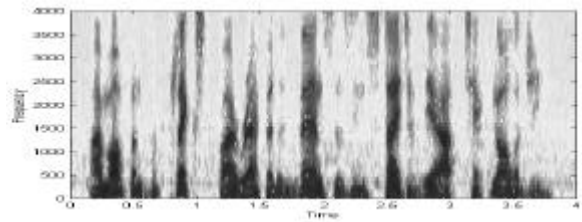
The proposed noise estimation algorithm was also applied to speech denoising using the perceptual Wiener filter. The spectrograms of the clean, noisy and denoised sentences are shown Fig. 3(a), 3(b) and 3(c), respectively. The noisy sentence is a short speech sentence mixed with car noise at a signal-to-noise ratio of about 5dB. Car noise is also a slowly changing non-stationary coloured noise. The noise is estimated on a frame basis using the proposed estimation algorithm. The estimate is then used to calculate the denoising gain (Eq. 3). Informal listening demonstrates that the denoised sentence is natural sounding without any detectable musical noise.



(a) Clean speech



(b) Noisy speech



(c) Denoised speech

Figure 3: Spectrograms of the clean, noisy and denoised sentences.

5. CONCLUSION

A fast and robust noise estimation algorithm has been proposed. The noise estimation is based on adaptive smoothing of the noisy signal power, with the smoothing parameter controlled by the estimated subband *a posteriori* signal-to-noise ratio. It produces accurate results even at very low signal-to-noise ratios.

The proposed noise estimation algorithm is designed to rapidly track non-stationary noise in the presence of speech, and does not depend on preliminary voice activity detection. Since it was developed for a subband system, it is also inherently suitable for tracking coloured as well as non-stationary noise.

The proposed noise estimation algorithm was developed especially for the perceptual Wiener filtering method of speech denoising [7, 9]. However, it could equally be used in virtually all other noise reduction algorithms, since most methods depend on having an accurate estimate of the noise energy.

Enhanced speech of good perceptual quality is produced by the perceptual Wiener filter using this noise estimation technique. Although we have used a critical band filterbank as a time-frequency decomposition tool, the proposed noise estimation and speech enhancement technique can also be extended easily to other filterbanks, e.g. those based on the DFT or the DCT.

Acknowledgement: This project is partially supported by a URSP-2001 Grant, Australia.

6. REFERENCES

[1] Ambikairajah, E., Epps, J. and Lin, L., "Wideband speech and audio coding using Gammatone filter banks," *Proc. ICASSP*, Salt Lake City, pp. 773-776, 2001.

[2] Brandenburg, K.B. and Stoll, G., "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780-792, 1994.

[3] Doblinger, G., "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Proc. Eurospeech*, Madrid, pp. 1513-1516, 1995.

[4] Gustafsson, S., Jax, P. and Vary, P., "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," *Proc. ICASSP*, Seattle, pp. 397-400, 1998.

[5] Lim, J.S. and Oppenheim, A.V., "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, 1979.

[6] Lin, L., Ambikairajah, E. and Holmes, W.H., "Auditory filterbank design using masking curves," *Proc. Eurospeech*, Aalborg, pp. 411-414, 2001.

[7] Lin, L., Holmes, W.H. and Ambikairajah, E., "Speech enhancement based on a perceptual modification of Wiener filtering" *Proc. ICSLP*, Denver, pp. 781-784, 2002.

[8] Lin, L., Holmes, W.H. and Ambikairajah, E., "Speech denoising using a perceptual modification of Wiener filtering," accepted for *Electronics Letters*, 2002.

[9] Martin, R., "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, 2001.

[10] Vaseghi, S.V., *Advanced Digital Signal Processing and Noise Reduction*. NY: John Wiley, 2nd ed., 2000.

[11] Virag, N., "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126-137, 1999.