SUB–BAND ADAPTIVE FILTERING APPLIED TO SPEECH ENHANCEMENT

David J. Darlington, Douglas R. Campbell

Department of Electrical and Electronic Engineering, University of Paisley,
High Street, Paisley PA1 2BE, Scotland, UNITED KINGDOM.

Tel: **+44 141 848 3428** *Fax*: **+44 141 848 3404**
E–mail: darl_ee0@helios21.paisley.ac.uk

ABSTRACT

An adaptive noise cancellation scheme for speech processing is proposed. In this, the adaptive filters are implemented in frequency–limited sub–bands. In previous work, the filters had been distributed in a linear fashion in the frequency domain. This work investigates the effects of spacing the filters more in sympathy with the signal power and spectral characteristics. It emerges that improvements in signal–to–noise ratio of processed noisy speech signals may be obtained when the sub–bands are spaced according to a published cochlear function.

1. INTRODUCTION

1.1 Speech enhancement system

A multi–channel, sub–band adaptive system for enhancement of speech signals corrupted by background noise is being developed by the authors. Enhancement in this context means improvement of the quality or intelligibility of the speech signal, by reduction of background noise or speech distortion, and hence improvement of the signal to noise ratio of the contaminated speech. This may be desired to render the speech more intelligible either to a human listener (e.g. in a hearing aid system or hands–free mobile telephone), or an automatic speech recogniser.

1.2 Sub–band scheme based on cochlear model

A sub–band system decomposes the wideband input signals into a number of band–limited signals, superficially similar to the treatment the human ear performs on incoming signals. A significant advantage of using sub–band processing for speech enhancement is that it allows for diverse processing in each sub–band depending on factors such as signal power, noise power and level of coherence between signal and noise in the two channels. For instance, if it can be determined that one particular sub–band contains no speech information, that particular band could simply be blocked from passing through the processor. Alternatively, if a particular band contains no noise energy, it is possible that processing would actually degrade the signal unnecessarily and so this band could be passed unchanged.

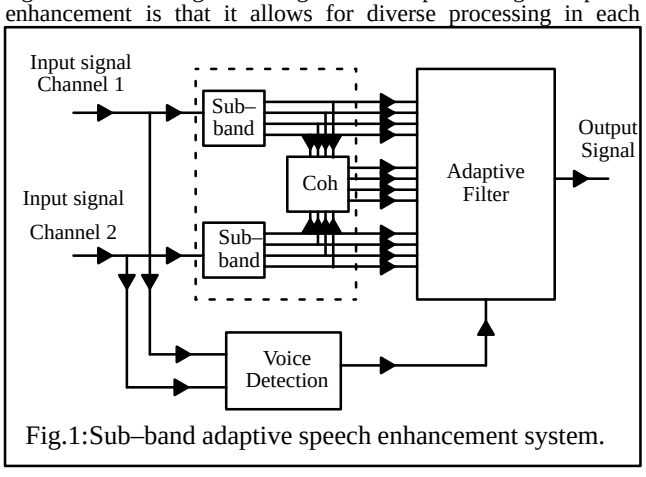

Fig.1:Sub–band adaptive speech enhancement system.

In addition, the implementation of a classical adaptive noise cancellation scheme in a number of frequency–limited sub–bands permits faster convergence of the filter coefficients due to the reduction of signal power and adaptive filter length in each sub–band. The system under development is shown in Fig.1 below.

2. COCHLEAR MODELLING

2.1 Human hearing

In speech enhancement research, new ideas are often stimulated by study of auditory processes, since humans are capable of detecting and understanding speech at low signal–to–noise ratios without prior knowledge of the speech, the noise or the environment [1], [2]. An important and much–studied feature is that of the filterbank present in the cochlea, within the ear, which splits incoming signals into a large number of band–limited signals prior to further processing.

This work is specifically oriented towards enhancement of speech signals. The human cochlea evolved to deal wth all sounds available to the human ear. This paper forms part of an investigation into incorporating an engineering model of the cochlea into a multi–band adaptive noise cancellation scheme and determining whether it is advantageous to arrange the sub–bands with an ‘equalised power’ distribution.

2.2 Cochlear function

In the work of Toner and Campbell, the sub–band filters were linearly distributed in the frequency domain. However, for the case of modelling human hearing systems, this is not an accurate model of the cochlea. Ghitza [1] used a logarithmic function to approximate the distribution of filters in the human cochlea. However, Greenwood [3] has presented the following, more accurate function for the spacing of the filters in the mammalian cochlea:

$$F(x) = A\,(10^{ax} - k)\ \text{Hz} \qquad (1)$$

where x is the proportional distance from 0 to 1 along the cochlear membrane, A, a and k are constants based on empirical knowledge of the mammalian cochlea and $F(x)$ the upper and lower cut–off

frequencies for each filter obtained by the limiting values of *x*. For the human cochlea, values of A=165.4, a=2.1 and k=0.88 are used, and this is confirmed by Allen [4]. The number of filters within the cochlear filterbank is not accurately known, and different sources suggest various numbers of filters within their models. For instance, Allen [4] suggests 20 filters, Cooke [5] suggests 65 and Ghitza [1], [6] suggests 85 and 190. Sub–band adaptive filtering work has indicated that the greater the number of sub–bands used, the faster will be the convergence of the overall adaptive system. In this work, the filtering is achieved by modifying the spectra of the FFT of the input signals, and the number of filters is therefore limited by the size of the FFT.

3. ADAPTIVE NOISE CANCELLATION

The sub–band speech enhancement scheme described here is an extension of that of Toner and Campbell [8] and Campbell [9]. It uses the least mean squares (LMS) algorithm [10] in an adaptive noise cancellation (ANC) scheme [11]**,** to model the differential transfer function between noise signals in a number of sub–bands. In the classical noise canceller, shown in Fig.2 below, it is assumed that desired speech (***s***) is present only in one of the two channels (the primary) and that the noise signal ($\boldsymbol{n_2}$) at the reference input is highly correlated with the noise signal ($\boldsymbol{n_1}$) at the primary. The adaptive filter weights will converge to the differential transfer function between the two inputs, resulting in filter output ***y*** being an estimate of only the noise present in signal ***d***. The output ***e*** will therefore be an estimate of speech signal ***s***.

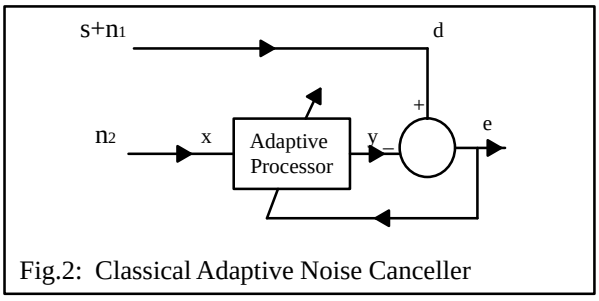

Fig.2: Classical Adaptive Noise Canceller

The multi–band approach reduces the problem of identifying a single, lengthy impulse response to one of identifying a set of shorter, parallel filters with approximately the same computational complexity as conventional LMS [7], [8]. Toner and Campbell [8] reported that the multi–band ANC approach considerably improved the initial mean square error (MSE) convergence rate. The improvement in algorithm convergence has been investigated by Mahalanobis et al [7] and the authors. Several aspects of this algorithm remain open for further investigation, such as the reasons for the improvement in algorithm convergence, the appropriate number of bands used in the decomposition, the spacing of these bands and the filter length in each sub–band.

Recent work by the authors and others has demonstrated that the improvement in convergence speed effected using a sub–band approach to adaptive filtering results from the larger stepsize permitted by the lower power and shorter filters used in each sub–band [7]. If the signals in each band are of the same power, and each band uses the same filter length, then each will converge at the same rate. However, if the signal in one band is of greater power, the stepsize within that band will be reduced and convergence within that band will be slow. Additionally, overall system convergence will be dominated by that band. Therefore, the convergence of the system should theoretically be optimised if the sub–bands are distributed such that the signals in each band are of equal power.

4. MSE CONVERGENCE

4.1 Experimental Setup

Speech shaped noise signals were obtained by weighting the frequency bins of a white noise signal in accordance with the universal long–term average speech spectrum (LTASS) determined by Byrne et al [12]**.** Speech–shaped and white noise signals, both band–limited to below 5kHz, were processed by an LMS–based sub–band adaptive noise canceller using three different sub–band filter spacing methods (linear, logarithmic and cochlear distribution) and four different numbers of sub–bands (1, 4, 16 and 64). The overall MSE convergence curves were obtained by summation over the frequency bands, which is justified by the verified orthogonal nature of the sub–band signals. For each test the results of 50 experimental runs were ensemble averaged, producing the log MSE plots of Fig.3.

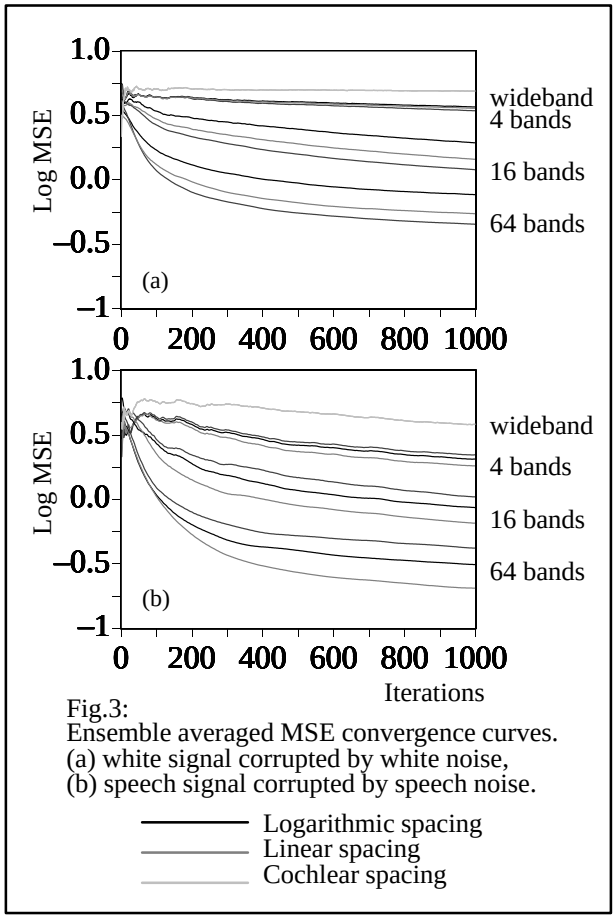

Fig.3:
Ensemble averaged MSE convergence curves.
(a) white signal corrupted by white noise,
(b) speech signal corrupted by speech noise.

4.2 Results

For the first test, independent white noise sequences were used to represent both signal and noise i.e. *s* and *n* of Fig.2. 'Signal' and 'noise' are each of unity variance. Fig.3(a) confirms that increasing the number of sub–bands improves the speed of convergence and lowers the steady–state MSE value. It also shows that for all sets of sub–bands used, the MSE converges to a lower steady–state value when linearly distributed filters are used. These equalise the signal

and noise power across the sub–bands.

The second test used uncorrelated speech–shaped noises as both *s* and *n*. Fig.3(b) shows that for all sets of sub–bands, the MSE reaches a lower steady–state value using the cochlear distribution. It has been verified that for this signal case, the cochlear spacing provides the closest approximation to equalisation of signal and noise power, followed by the logarithmic distribution which could be considered an approximation to cochlear spacing.

Tests were also performed using white noise as 'signal' and speech noise as 'noise', and vice versa. As might be expected, in these cases, the results were less clear cut than those of the preceding two sets of tests, with no particular spacing method appearing advantageous. It appears that there are ground for selecting sub–band filter spacing on the basis of known signal and noise characteristics, but that this will be of advantage only when these characteristics are stationary. In a more general case there may be grounds for selecting the sub–band distribution either for the worst case or most common case signal and noise scenario. The speech enhancement problem is not so simple a case as a speech shaped noise corrupted by either a white noise or an interfering speech shaped noise, as both speech and noise may be random and non–stationary. Also, it has been established that increasing the number of bands improves the MSE convergence and steady–state characteristics, but not how this affects the quality of the processed speech.

5. SPEECH TESTS.

5.1 Experimental Setup

Early tests [13] had suggested that using cochlear spaced filters for the sub–banding resulted in an improvement in the output SNR of up to 3dB compared with the linear case, when speech was used as the signal. This improvement was observed using both white and speech shaped noise as the corrupting noise signal. The purpose of these more extensive tests was to determine whether the this effect would still occur when a more realistic transfer function was being modelled, and to study the effect on the SNR of varying the length of the adaptive filter.

A clean, anechoic speech signal, sampled at 10kHz and 40,000 samples long was used as the signal *s* of Fig.2. In this test, six corrupting noise signals were used; white and speech shaped noise at high, medium and low SNRs. A 256–point simulated room acoustic transfer function was generated, modelling an acoustic path difference between the primary and reference inputs. Although this is still an idealised case, its gives more of an idea of how such a system will respond in a realistic environment. Three experimental runs were performed. The first run utilised the classical adaptive noise canceller of Fig.2. For the second and third runs, a more realistic speech enhancement scenario was modelled by adding the clean speech to the reference input. The second run used an intermittent processing scheme, where the adaptation is paused in the presence of speech. The start and endpoints of the speech were manually labelled. The third run used an 'adapt and freeze' approach, wherein the filters were only permitted to adapt during the first 3000 (speech–free) iterations, as an approximation to the intermittent approach which avoids problems due to incorrect determination of speech pauses. Wideband filter lengths of 64, 128, 256, 512 and 1024 were used for these tests.

Fig.4 shows the output SNR improvements, after convergence of the filters to the desired values, in 16 sub–bands, using the classical ANC of the first run.

5.2 Results

5.2.1 Classical Noise Cancellation

Using white noise, the cochlear spacing gives an improved output SNR over linear spacing in the vast majority of cases. It can be seen from Fig.4 that when the input SNR is low, the cochlear spacing improves the SNR by 2 to 3dB more than the linear spacing. As the corrupting noise dominates the signal more at lower SNRs, the SNRI becomes greater (in some instances) when linear filters are used. For instance, at an input SNR of –0.1dB, the cochlear spacing gives a higher SNRI in 4 sub–bands at filter lengths of 256 – 1024, whereas the linear spacing gives the best SNR at shorter filter lengths. In 16 bands (at this input SNR), the linear spacing produces the best output SNR for all filter lengths except 1024.

Using speech shaped noise, the cochlear spacing always proves better than linear spacing, and as input SNR decreases, the difference between log and cochlear spacing performance becomes less marked. For instance, in 16 bands at an input SNR of 8.2dB, the cochlear spacing produces an SNRI improved by 3.5dB over either the linear or log cases. Whereas, at an input SNR of –10.9dB in 4 bands, the difference in SNRI between log and cochlear processing is negligible, but each is 2.5 to 3dB greater than the linear case.

5.2.2 Intermittent ANC

Results using the actual intermittent ANC were less conclusive, possibly due to problems with the manual labelling of the speech segments of the signal. However, the 'adapt and freeze' method is a good approximation to this.

At very high SNRs, the processing actually degraded the signal in some instances. However, cochlear spacing almost invariably proved best at improving the SNR, as in 5.2.1. Only for an SNR of –0.1dB, and filter lengths shorter than 256 (4–band case) or 1024 (16–band case), did the linear distribution perform better. Logarithmic filters actually showed an improvement over cochlear filters for speech shaped noise at high SNRs in 16 sub–bands, using filters of length 512 and 1024.

5.2.3 Complexity of filters

It is also the case that for most of these tests, the best SNRI was achieved using a wideband adaptive filter length of 256, split into filters of length 64 and 16 for the 4–band and 16–band case respectively. However, the results are not entirely conclusive here, although at all times the SNRI is better when the adaptive filter complexity is at least that of the transfer function being modelled.

6. CONCLUSIONS

It has been demonstrated previously [13] that using cochlear spaced sub–bands can improve the output SNR of speech signals in a classical, sub–band noise cancellation scheme with no additional distortion apparent. It is also evident that increasing the number of sub–bands improves convergence time and lowers the steady–state MSE. In the tests reported here, using a more realistic model of the cochlea gave improved results for all cases where the noise had a similar spectrum to speech. When the noise spectrum was not biased towards the low end (i.e. using white noise) then using a cochlear model produced better results in some cases, especially at higher input SNRs, whereas using linear spacing produced better results in some cases at lower SNRs.

Therefore, for applications involving enhancement of speech signals, if it is known that the noise power is low or that the noise spectrum is biased toward lower frequencies, the evidence suggests that cochlear spaced filters will give better noise reduction. Using filtering complexity of at least that of the acoustic transfer function

will also increase the SNR improvement. However, increasing the filter lengths beyond that which is necessary for modelling of the acoustic transfer function will slow filter convergence and (informal listening tests suggest) add distortion to the signal by nature of echoes. Investigation into this is ongoing, as is investigation into methods of quantifying the distortion introduced by the processing, to ascertain whether the increases in SNRI are compromised by a distortion of the actual speech.

7. REFERENCES

1. O. Ghitza. *Auditory models and human performance in tasks related to speech coding and speech recognition*. IEEE Transactions on Speech and Audio Processing **2** (1): 115–132, 1994.
2. Y. M. Cheng, D. O'Shaughnessy. *Speech enhancement based conceptually on auditory evidence*. IEEE Transactions on Signal Processing **39** (9): 1943–1954, 1991.
3. D. D. Greenwood. *A cochlear frequency–position function for several species – 29 years later.* Journal of the Acoustical Society of America **86** (6): 2592–2605, 1990.
4. J. B. Allen. *How do humans process and recognise speech?*. IEEE Transactions on Speech and Audio Processing **2** (4): 567–577, 1994.
5. M. P. Cooke. *NPL computer model of peripheral auditory processing*. NPL Report DITC 58/85, 1985.
6. O. Ghitza. *Auditory nerve representation as a front–end for speech recongition in a noisy environment*. Computer Speech and Language **1**: 109–130, 1993.
7. A. Mahalanobis, S. Song, S. K. Mitra and M. R. Petraglia. *Adaptive FIR Filters based on Structural Subband Decomposition for System Identification Problems*. IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing **40** (6): 375–381, 1993.
8. E. Toner and D. R. Campbell. *Speech Enhancement using sub–band intermittent adaption.* Speech Communication **12:** 253–259, 1993.
9. D. R. Campbell. *Adaptive speech enhancement with diverse sub–band processing*. EUSIPCO–94 Signal Processing VII: Theories and Applications September 13–16: 1210–1213, 1994.
10. B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice Hall, New Jersey, 1995.
11. B. Widrow, J. R. Glover jr., J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong jr. and R. C. Goodlin. *Adaptive Noise Cancelling: Principles and Applications*. Proceedings of the IEEE **63** (12): 1692–1716, 1975.
12. D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Heth, J. Kei, C. Lui, J. Kiessling, M. N. Kotby, N. H. A. Nasser, W. A. H. El Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, R. Meredith, T. Sirimanna, G. T. Kiladze, G. I. Frolenkov, S. Westerman and C. Ludvigsen. *An international comparison of long–term average speech spectra*. Journal of the Acoustical Society of America **96** (4): 2108–2120, 1994.
13. D. J. Darlington, D. R. Campbell, *Sub–band adaptive filtering applied to speech enhancement*, ESCA Tutorial and Research Workshop, The Auditory Basis of Speech Perception, Keele University 15–19 July 1996.

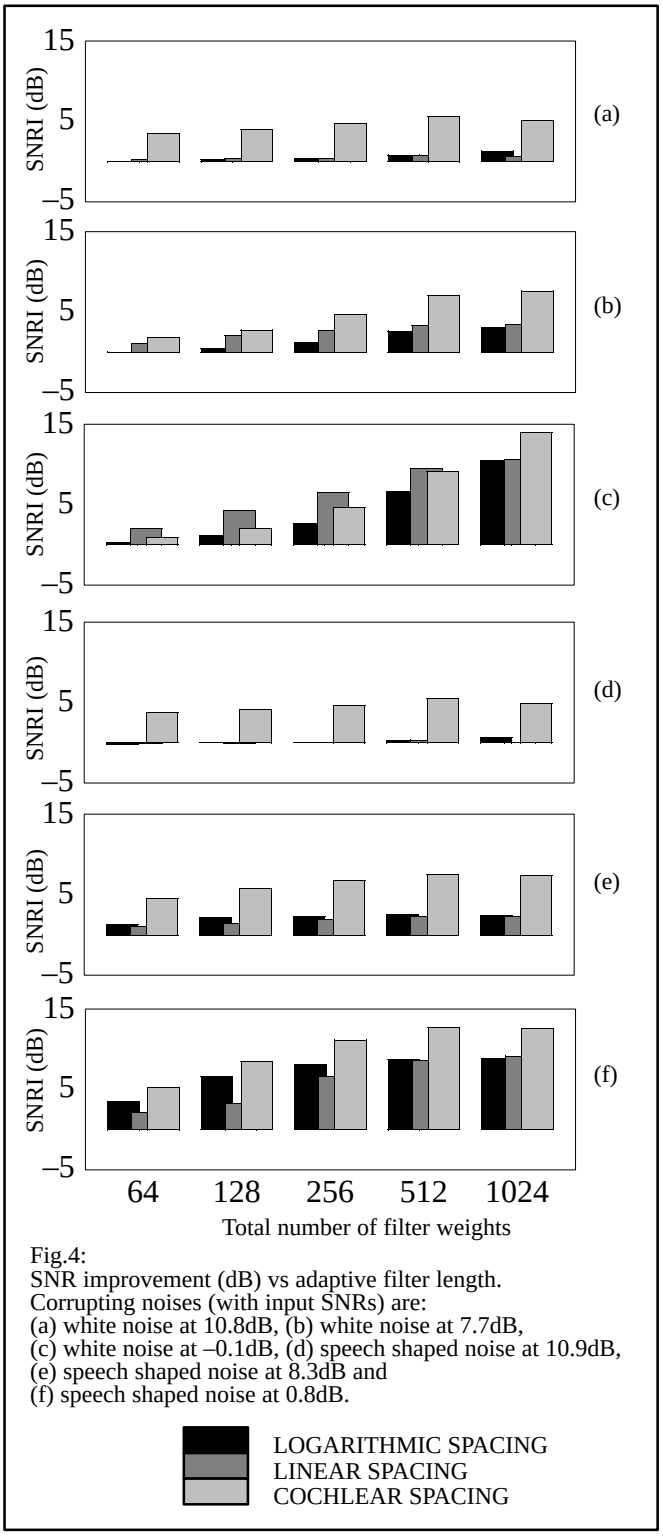

Fig.4:
SNR improvement (dB) vs adaptive filter length.
Corrupting noises (with input SNRs) are:
(a) white noise at 10.8dB, (b) white noise at 7.7dB,
(c) white noise at –0.1dB, (d) speech shaped noise at 10.9dB,
(e) speech shaped noise at 8.3dB and
(f) speech shaped noise at 0.8dB.