Adaptive Filtering using a Highly Oversampled Weighted Overlap-Add Filterbank in an Ultra Low-Power System

Robert L. Brennan[1], *Reza Abutalebi*[2], *Hamid Sheikhzadeh*[1]

[1]Dspfactory Ltd., 611 Kumpf Dr., Unit 200, Waterloo, Ontario, Canada N2V 1K8
e-mail: rob.brennan@dspfactory.com
[2]EE Dept., Amirkabir University of Technology, Tehran, Iran

ABSTRACT

Filterbank (multi-rate) analysis and synthesis strategies prove advantageous in many signal processing areas operating as a divide and conquer strategy tackling difficult problems into an equivalent series of much simpler problems. For example, large convolutional systems encountered in applications such as echo cancellation and feedback cancellation may require a large number of filter taps. Using the filterbank technique, it may equivalently be implemented as a parallel combination of much shorter subband filters. When properly designed, the filterbank subband signals are minimally overlapping in frequency yielding signals that are approximately orthogonal to each other. Lately, digital filterbank techniques, with their great precision, have enabled many strategies to be implemented that were difficult or impractical with analog structures. Accordingly, much theory has been developed including the so-called perfect reconstruction filterbank. An oversampled DFT filterbank using WOLA (weighted overlap-add) processing provides an extremely efficient and elegant solution. This paper will describe this filterbank within a dedicated ASIC and algorithmic procedures for casting many algorithms into a multi-rate framework.

1. Introduction

Numerous advantages are obtained using the multi-rate framework. Adaptive filtering techniques typified by the LMS algorithm is greatly affected by the eigenvalue spread problem. In short, the LMS algorithm converges slowly when the noise reference is non-white implying a large ratio between the maximum and the minimum correlation matrix eigenvalues. The subband approach represents the original spectrum as a parallel combination of much whiter subband signals. The inherent scaling of the subbands largely captures the original coloration.

Subbands may be adapted separately. This is a result of the near-orthogonality between bands. This allows great flexibility in adapting subband filters when processing power is limited. Again, because of subband orthogonality, each subband may be adapted with separate convergence factors. This is useful in applications where narrowband disturbances exist allowing processing power to be concentrated only on the affected bands.

The filtering operation complexity is greatly reduced by converting intensive time-domain convolutions to relatively short frequency-domain convolutions. In certain cases, the filtering in each parallel path may be reduced to multiplication by a single (possibly) complex value.

Adaptive filtering applications, the main subject of this paper are a different class of algorithms where the filterbank is not intended to reconstruct the input signal directly but after modifications have been made to the analysis signal. Typically, the filterbank is invoked to model a desired system (as in hearing aid applications) or to model an undesired or disturbance system to facilitate its cancellation (as in echo cancellation systems). These modifications range from scalar real or complex multiplications [4] as in hearing aid or vector complex multiplications in the case of adaptive filtering structures.

2. The DSP System

The single-chip DSP system [8] is implemented on a 0.18μ CMOS technology containing the mixed-signal portion, DSP core, RAM, the weighted overlap-add (WOLA) filterbank, and the input-output processor (IOP. A separate E^2PROM provides non-volatile storage. The RAM consists of two 4K-word data spaces and a 12K-word program memory space. Additional shared memory for the WOLA filterbank and the IOP is also provided. The core provides 1 MIPS/MHz operation and has a maximum clock rate of 4 MHz at 1 volt. The WOLA processor provides approximately 4 MIPS/MHz. At 1.8 volts, 30 MHz operation is also possible. The entire system operates on a single battery down to 0.9 volts and consumes less than 1 mW. This chipset is packaged into a

0-7803-7576-9/02$17.00 © 2002 IEEE

6.5 x 3.5 x 2.5 mm hybrid circuit. The DSP core communicates with the outside world through a UART (serial port), 16 general-purpose input/output pins and a channel dedicated to the samples coming from and going to the mixed-signal chip.

3. Wola Architecture

The WOLA coprocessor performs M channel complex subband decomposition [5-7] on a microcoded architecture using dedicated hardware resources for flexibility and high efficiency. The main operations will be now be detailed.

On input, the WOLA coprocessor forms L sample frames by concatenating a number of R -sample segments together advancing one segment each analysis [1]. Each frame, therefore, overlaps the previous frame by $L-R$ samples.

During analysis (Figure 1), each L -sample frame is multiplied point-by-point by a prototype low-pass filter (window) $W(1:L)$ (note the use of Matlab notation) and folded (time-aliased) into N -sample blocks, $Z(1:N)$ ready for (evenly-stacked or oddly-stacked) FFT modulation. The use of complex modulation produces $M=N/2$ unique bands when the odd FFT is used and $M=N/2+1$ bands when the even FFT is used (Figure 3). The input signal must be flipped in sign every N samples by the *sign sequencer* block when odd-stacking is used. This operation is undone in the synthesis stage.

For computational efficiency, L and R and N are all powers-of-two; L is additionally related to R and N as a power-of-2 multiple. Similarly, N is a power-of-two multiple of R.

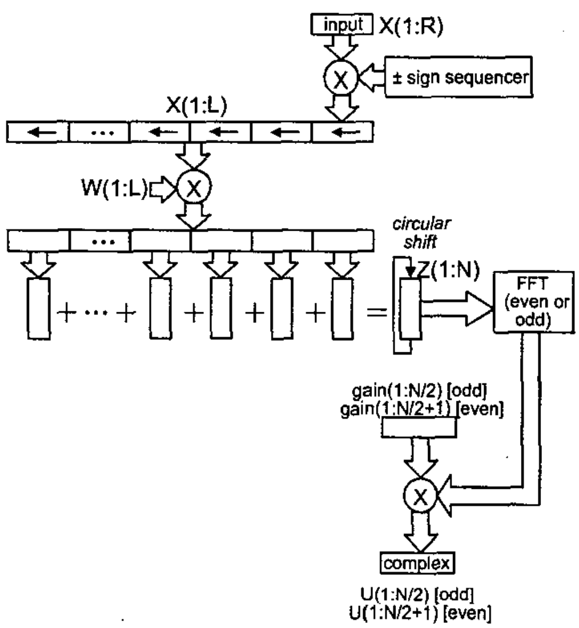

Figure 1. – WOLA Analysis

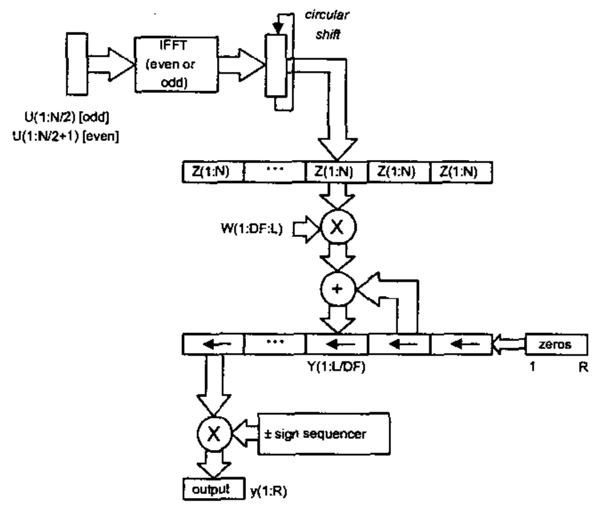

Figure 2. – WOLA Synthesis

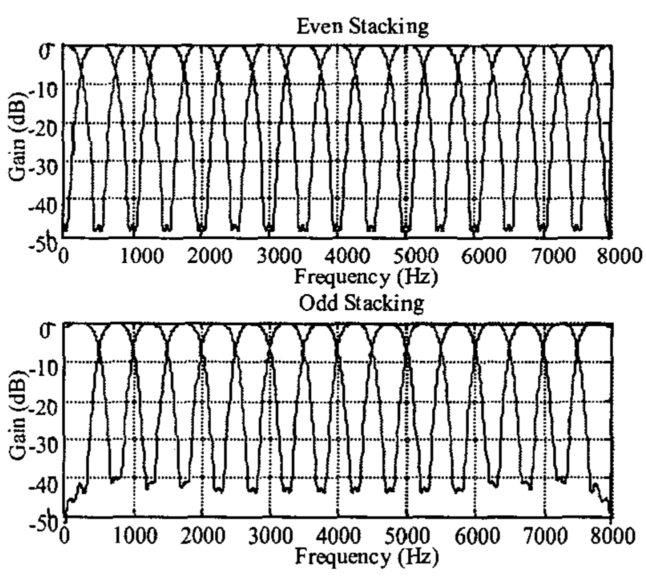

Figure 3. – Filterbank bands for even/odd stacking

During synthesis (Figure 2), L' (possibly shorter than L to reduce delay) sample overlapping blocks are produced from N sample inverse FFT'ed blocks through periodic repetition weighted by the synthesis window W' and overlap-added to the output FIFO.

For adaptive filtering applications, the WOLA is operated in stereo analysis mode. In this mode, the WOLA processes two simultaneous data streams, combined as a complex signal, and performs a final butterfly separation step to produce two real transforms. Transformations are applied to the separated stereo channels by the DSP core, which are then mixed together in the frequency domain. This mixed signal is then returned to the time domain via a WOLA synthesis transformation

4. Adaptive Filtering

Adaptive algorithms are well-known and extensively used for interference cancellation systems. The interference is modeled as an additive disturbance to the desired signal after passing through a slowly time-varying system (disturbance model). The goal of the adaptive algorithm is to cancel the disturbance from the noisy input signal by generating a counteracting influence using information gathered by observing the disturbance path dynamics. In most acoustic applications, a linear model suffices. Furthermore, if the model is restricted to an FIR system, well known algorithms such as the LMS or RLS algorithms may be efficiently employed to *track* the disturbance model.

Although many more complex adaptation techniques may be used, the LMS algorithm is extensively used for its stability, good performance and relatively low computational complexity. Nevertheless, these advantages tend to dissipate as increasingly complex (but realistic) systems are modeled. Echo cancellation, for example, may require thousands of taps. It is highly desirable to decouple difficult large problems such as these into a combination of simpler parallel problems.

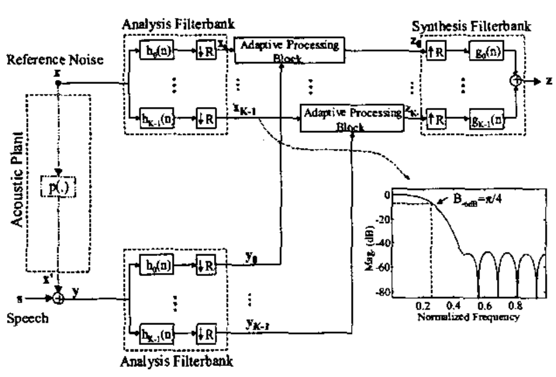

Figure 4. Block diagram of the SAF system.

Subband adaptive filtering (SAF) (figures 4 and 5) provides an efficient means for the decoupling just mentioned. As the number, M, of subbands increases, the global spectral dynamics are increasingly approximated as a sequence of relatively flat subbands accompanied by scalar gain – a *stair-step* approximation. As a result, fewer taps per subband are necessary to account for the residual deviation from flatness.

Specifically, the convergence of the fullband LMS algorithm and variants is reduced when there is a large eigenvalue spread in the reference correlation matrix occurring when there is significant coloration in the input spectrum. The update formula for adaptive filter $\mathbf{w}_n$ is given as $\mathbf{w}_{n+1} = \mathbf{w}_n + \mu\, z_n \mathbf{x}_n$ where μ is the step size and z_n is the enhanced output $z_n = y_n - \mathbf{w}_n^T \mathbf{x}_n$

The MSE, ξ_n, for large values of M (16 or larger) may be approximated [8] by a fixed portion, $\xi_{\min}$, and a transient portion which is a function of the eigenvalues of the reference correlation matrix. $\xi_{\min}$, is the theoretical minimum misadjustment at convergence given by:

$$\xi_{\min} = E\{s_n^2\} + E\{x_n'^2\} - \mathbf{r}_{\mathbf{xx}'}^T \mathbf{R}_{\mathbf{xx}}^{-1} \mathbf{r}_{\mathbf{xx}} \tag{1}$$

where $\mathbf{R}_{\mathbf{xx}} = E\{s_n^2\}$ and $\mathbf{r}_{\mathbf{xx}'} = E\{x_n x_n'\}$ are the correlation matrix of the reference noise and the cross-correlation vector between the vector reference input, $\mathbf{x}_n$, and the scalar filtered reference, x_n'. To normalize the results by removing any signal power influences, a unity power assumption is made (diagonal entries of the correlation matrix represent the signal power) leading to:

$$tr(\mathbf{R}_{\mathbf{xx}}) = M\,. \tag{2}$$

Given these definitions, it can be shown by Morgan [9] that a modal power approximation of the is given by:

$$\xi_n = \xi_{\min} + \frac{1}{M}\sum_{m=0}^{M-1} \lambda_m (1 - \mu\lambda_m)^{2n} \tag{3}$$

The second term above governs the transient portion and, therefore, the convergence rate. As can be seen, the form of the summand above suggests that the λ_m term and the $(1-\mu\lambda_m)^{2n}$ term have competitive influences. It is possible to determine the dominant eigenvalue λ_k at time n_0 from the given set of M eignevalues by differentiating the equation $f(\lambda) = \lambda(1-\mu\lambda)^{2n}$ with respect to λ, equating to zero and selecting the closest eigenvalue. This differentiation yields, $\lambda = \frac{1}{\mu(1+n_0)}$

As time index, n_0 advances, it is clear that, the largest eigenvalues dominate the early convergence while the smallest eigenvalues dominate the latter convergence. The goal of the WOLA filterbank, as described earlier, is to flatten the dynamic range of the eignevalues, λ_m which has been shown to speed convergence dramatically. If critical sampling is used, aliasing distortion manifests itself from spectral leakage between adjacent bands. It is highly impractical to eliminate this distortion through higher order filters because the required order is extremely high. Instead aliasing is either eliminated through adaptive cross-filters or gap filters [2-3]. These systems are generally cumbersome requiring higher numbers of computations and in the case of gap filter systems, produce audible distortion.

Oversampled systems relax the requirements on aliasing rejection by not using the maximum amount of down-sampling possible. Highly oversampled systems (oversampling ratios of $OS = 2$ or more) are of particular interest because they provide a framework upon which large modifications can be made to signals (in the signal-path) without significant distortion. For practical purposes, it is desirable that the filterbank control both signal-path and control-path functionality. In the case of adaptive filtering, the filterbank should provide suitable inputs for the adaptive filter update block – a control-path function while performing the filtering in the signal-path (figure 5). Applications include digital hearing aids, echo cancellation systems and feedback cancellation systems.

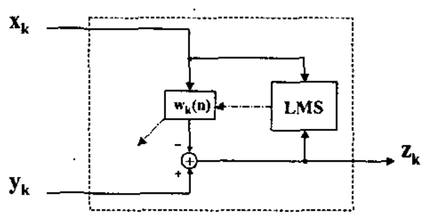

Figure 5. Adaptive processing block for each subband.

Although oversampling by its nature sidesteps the aliasing problem, it leads to a new problem – each band is only partially full, occupying only the low frequency region up to π / OS where the oversampling ratio, OS, is given by N / R. The remaining frequency region from π / OS to π is suppressed by the analysis filter (figure 1). Thus, the action of oversampling, while simplifying gain adjustments in the signal path, leads to a convergence rate reduction for control-path functions just as in the fullband case. Fortunately, the non-white nature of the subband signal is a fixed and non-time varying low-pass filter. This allows the proposition of effective techniques to ameliorate this problem [10]

In the *Whitening-by-Decimation* (WBD) technique, it is observed that, if the convergence problem is caused by oversampling then decimation might be employed, in the control-path (for adaptive filter calculation purposes), to re-expand the subband signal to a greater proportion of the total subband spectrum. It is important that this decimation not be carried out fully since the aliasing problem will reappear leading to misadjustment of the LMS parameters.

It was experimentally determined that the suitable final effective oversampling rate should be two. Therefore, of course, this technique is applicable only for systems with oversampling ratios higher than two. Since the control-path calculations are performed at a decimated rate, it is necessary to interpolate the calculated filter back to the original oversampling factor for filtering in the signal-path. This interpolation may be done by simply inserting zeros since the generated images fall in frequency regions already suppressed by the analysis filter. Moreover, these zeroed coefficients represent multiplies that may be pruned leading to a reduction in computational effort.

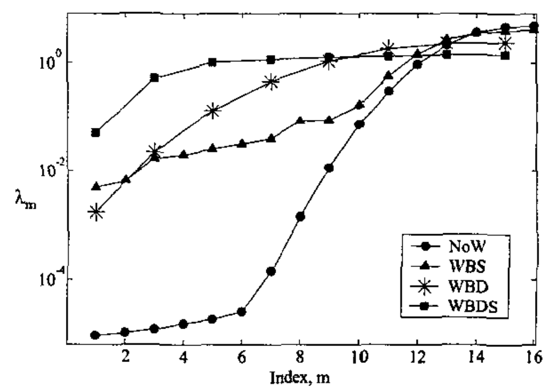

Figure 6. Eigenvalues of the input correlation matrix in the NoW, WBS, WBD and WBDS cases.

In the *Whitening-by-Spectral-Emphasis* (WBS) technique, an alternative technique is equalization of the subband spectrum to compensate for the suppressed high spectral region. This whitening is performed only in the control-path (as in the WBD method) both for the noisy signal path and the noise reference signals. Since the same filter is applied to both inputs to the LMS block, the modified algorithm still converges to the same solution except at an accelerated pace. In practice, for an original oversampling factor of four, a fourth order emphasis filter is necessary for satisfactory results.

5. Discussion

Combining the WBS and WBD techniques yields a very powerful system since they have complementary effects. The WBS system is most effective when the oversampling ratio is small since the oversampled subband signal occupies a relatively wide spectral region and only a small boost is necessary. The WBD technique works best when the oversampling is large and works best as long as the total oversampling factor remains greater or equal to two.

'Clearly these effects can be combined in the *Whitening-by-Decimation-and-Spectral-Emphasis* method; the WBD system first brings down the initial oversampling factor down to two leaving the WBS system to provide the rest of the effort by performing a small boost. In practice, the WBS system now only requires a second order filter – a considerable savings over the fourth order filter employed when used in the four times oversampled system. Figure 6 shows the relative improvement, note that NoW refers to the original uncompensated system with *No-Weighting*. Since the sum of the eigennvalues is fixed (2), increasing the smallest eigenvalues necessarily reduces the large ones. Although the initial convergence is slightly slowed as a result, the most important final convergence region is dramatically accelerated.

Conclusions

Adaptive filtering has been discussed in the context of a highly oversampled subband system. It was shown that correlation in the noise reference matrix results in slow convergence when the fullband LMS algorithm is applied. Critically sampled subband algorithms show great promise in solving this problem but require additional processing to compensate for the generated aliasing distortion. Highly oversampled subband systems sidestep this last problem but, without compensation, converge slowly because of the high coloration due to the oversampling itself. Two methods, *Whitening-by-Decimation* and *Whitening-by-Spectral Emphasis* provide straightforward yet effective means to overcome this limitation. Moreover, combining these methods yields a powerful adaptive system with all the advantages of subband processing and greatly accelerated convergence.

REFERENCES

[1] R. Brennan and T. Schneider, "A Flexible Filterbank Structure for Extensive Signal Manipulations in Digital Hearing Aids", Proc. IEEE Int. Symp. Circuits and Systems, pp.569-572, 1998.

[2] A. Gilloire, A. and Vetterli, M, "Adaptive filtering in subbands with critical sampling: analysis, experiments and application to acoustic echo cancellation," *IEEE Tran. Signal Processing*, vol. SP-40, no. 8, pp. 1862-1875, Aug. 1992.

[3] Shynk, J.J, "Frequency-domain and multirate adaptive filtering," *IEEE Signal Processing Magazine*, pp. 14-37, Jan. 1992.

[4] Schneider, T., Brennan, R.L., "A Multichannel Compression Strategy for a Digital Hearing Aid,". Proc. ICASSP-97, Munich, Germany, pp. 411-415

[5] Vaidyanathan, P.P., *Multirate Digital signal Processing*. Prentice-Hall Inc., 1993.

[6] Crochiere, R.E. and Rabiner, L.R., *Multirate Digital Signal Processing*. Prentice-Hall Inc., 1983.

[7] P.N. Heller, T. Karp and T.Q. Nguyen, "A General Formulation of Modulated Filter Banks", IEEE Trans Signal Processing, pp.986-1002, 1999.

[8] Brennan, R.L. et al., "An ultra low-power, miniature speech CODEC for 8 kbs and 16 kbs", Proc. ICSPAT 2000. Dallas, TX, 2000.

[9] D. R. Morgan, "Slow asymptotic convergence of LMS acoustic echo cancelers", *IEEE Tran. on Speech and Audio Processing* , vol.3, no.2, pp. 126-136, Mar. 1995.

[10] H. R. Abutalebi, H. Sheikhzadeh, R. L. Brennan, and G. H. Freeman, "Whitening Methods for Convergence Improvement of Highly Oversampled Subband Adaptive Filters", Submitted to ICASSP-2003