Ultra Low-Power Noise Reduction Strategies Using a Configurable Weighted Overlap-Add Coprocessor

R. Brennan, T. Schneider, W. Zhang
Dspfactory Ltd
611 Kumpf Drive, Unit 200
Waterloo, Ontario, N2V 1K8, Canada

Abstract

The availability of deep Sub-micron technology opens the door to advanced noise reduction algorithms specifically targeted for ultra low-power portable applications like hearing aids. These applications are extremely constrained by small physical size and extremely low power consumption requirements. The Weighted Overlap-Add (WOLA) filterbank discussed here provides a powerful platform for the implementation of noise reduction algorithms.

Introduction

This paper presents two extremely low-power noise reduction systems suitable for the demanding application of hearing aids and for industrial applications where extremely low power and extremely small size are required (less than 1mA at 1V and 17 square mm). Key to successful implementation in these areas is a tight fit between hardware architecture and the algorithms.

Most signal processing algorithms can be cast efficiently into a frequency-domain-processing framework. Since noise and speech are time varying frequency dependent quantities, noise reduction naturally fits in this framework. It is rare that noise reduction alone is the end result of signal processing. One typical application is processing for hearing loss (hearing aids). Other applications that work efficiently in conjunction with noise reduction include dynamic range compression, sub-band coding, directional processing, voice activity detection and echo cancellation. For these types of real-time audio signal processing applications, the filtering requirements are strict: i) low group delay, ii) high degree of adjustability, and iii) high fidelity. A frequency domain approach is an efficient method of meeting these constraints while delivering low power and flexibility.

This paper describes two types of noise reduction algorithms: i) a robust extremely low-power spectral subtraction noise reduction algorithm and ii) a low-delay noise attenuation algorithm more suited for digital hearing aid applications. Both algorithms are tightly coupled with a highly optimized, extremely low-power WOLA (Weighted Overlap-Add) filterbank [1].

The incoming speech is digitized at a sampling rate of 16kHz, presented to the analysis filterbank in overlapping blocks and split into a programmable number of uniform bands (frequency domain). After processing in the frequency domain, the frequency bands are combined together in the synthesis filterbank to produce time-domain-overlapping blocks. These blocks are weighted and summed together to produce the processed outgoing speech.

The choice of the number of bands, from 4 to 128, depends on the application. For hearing aid applications, 16 bands (500Hz each) or 32 bands (250 Hz each) provide excellent frequency selectivity and low delay (6 ms and 12 ms respectively). Non-uniform channels are created from the uniform bands through grouping. Hearing aid processing, to compensate for the hearing loss, occurs in the frequency domain (i.e. directly in the frequency bands). Noise reduction may be added to the hearing loss compensation directly in this channel structure.

Two extremely low power, small size noise reduction systems are presented in this paper: One for hearing aid applications and another for industrial applications. These algorithms have been tightly integrated with an extremely low-power WOLA filterbank to achieve extremely low system power consumption and small size [2].

WOLA Coprocessor

The WOLA design provides a highly flexible time-frequency representation amenable to sub-band adaptive algorithms, sub-coding and other similar applications [1], [2], [4]. The co-processor interfaces to a DSP core via shared memory (RAM).

The co-processor has two main sub-blocks (**Figure 1**): the WOLA and the Input/Output processor (IOP). Input samples are stored in a circular input FIFO. Every R (input block size) samples a WOLA analysis transformation is performed on L samples (L >> R).

In the noise reduction applications, the core is primarily used to analyze the incoming spectrum and to apply, via the shared RAM, appropriate attenuation factors for each frequency band. Then, the WOLA coprocessor performs a WOLA synthesis transformation and stores the results in the output FIFO. The

IOP is responsible for interpolating outgoing samples and decimating incoming samples.

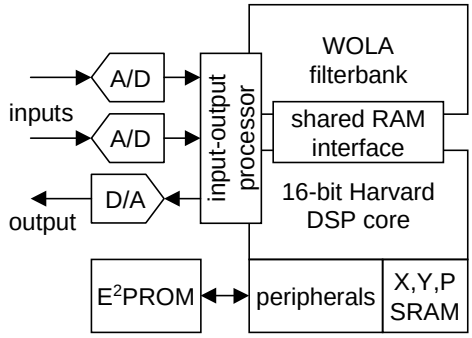

Figure 1: Overview of the co-processor's environment

Spectral Subtraction Noise Reduction

Crucial to this algorithm is the generation of a noise spectral estimate. During speech pauses, the noise spectral estimate is updated from the input (since it contains noise only) using long time averaging (about 1-2 sec). A voiced/unvoiced detector is used to determine the gaps in speech. In addition to the modulation technique described later, the narrow-band structure allows other possibilities including the detection of the pitch fundamental frequency. This fine spectrum structure is not visible in the coarser frequency structure mentioned in the Low Delay Noise Attenuation section. Once the noise estimate is known, it is used to calculate a frequency domain filter to suppress the background noise.

For stringent applications where a separation between speech and noise is required, a filterbank with narrower channels must be used. In these applications, the WOLA filterbank is configured to provide 128 narrow bands (62.5Hz each). In the algorithm that will be presented, these bands are grouped into 24 channels approximating Bark frequency spacing (**Figure 2**).

Given knowledge of the noise spectrum in each channel, enhanced speech is generated by a form of subtraction between the incoming noisy spectrum and the noise estimate. Since the noise and speech are divided into narrow bands, it is possible to affect a separation by preserving stronger speech components while suppressing nearby (in frequency) noise.

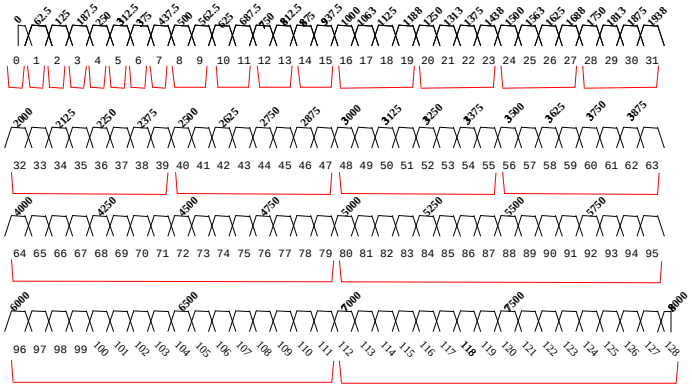

Figure 2: Band grouping approximating Bark frequency spacing

The flexible WOLA structure is easily adapted to a 128 band filterbank (narrow bands of 62.5 Hz each). In this mode, the WOLA performs FFT transformations on overlapping 256 point sine windowed input sample sections with 50 percent overlap. After spectral modification, the results are again weighted by a sine window and overlap-added.

To reduce computation while remaining faithful to the human auditory system, the number of bands was reduced to 24 (approximately) Bark spaced frequency *channels* (**Figure 2**). This brings the computation of gains down from 128 to 24 for savings of about 5 times.

Algorithm

Since the clean speech is not known, the optimal attenuation function, $H(\omega)$, must be estimated from the corrupted speech. $X(\omega) = S(\omega) + N(\omega)$.

The final update equation is given by:

$$\hat{H}(\omega) = \left\{ \frac{|X(\omega)|^2 - \beta |N(\omega)|^2}{|X(\omega)|^2} \right\}^{\alpha}$$

This formula is quite general and includes the Wiener (minimum square error) solution if β and α are set to 1.0 and 2.0 respectively. Parameter α controls how fast attenuation increases as SNR decreases. A value of 1.0 was used. Parameter β is the so-called over-subtraction factor. Residual noise and perceptual quality can be increased by setting β to values greater than 1.0.

Since the quantities used to calculate $H(\omega)$ are estimates, negative values can result from inaccuracies. To avoid this problem, a spectral floor (minimum value) for $H(\omega)$ is used. A value of –30 dB was used.

Complexity Reduction

For extremely low-power systems, the algorithmic complexity must be minimized. Often, this minimization can be done with little or no perceptual degradation. Already, one technique was mentioned, the grouping of bands into Bark spaced channels. While this procedure saves power, it actually reduces the musical noise artifact common in these systems by essentially *averaging* a number of frequency bands together.

It is advantageous to recast the previous attenuation equation into dB (assume that α is 1).

$$\hat{H}(\omega) = 1 - \frac{\beta |N(\omega)|^2}{|X(\omega)|^2}$$
$$= 1 - \beta \log_{10}^{-1}(-(X_{dB}(\omega) - N_{dB}(\omega))/10)$$

The availability of fast math libraries including logarithmic and exponential functions enable quick conversion to and from dB is simple. Aside from the antilog required, this formula is much simpler when $X(\omega)$ and $N(\omega)$ are kept track of in dB. In fact, $X_{dB}(\omega) - N_{dB}(\omega)$, is just the SNR at a particular frequency.

Time-Slicing

A considerable reduction in computation is achieved by reducing the update rate for the noise attenuation parameters. Time slicing operates over 4 time slots. Eight frequency channels are computed at a time during the first three; the last slot is reserved for the voice activity detection algorithm.

Since only a selected number of channels are updated every pass through the algorithm, overhead is created because the algorithm must keep track of the partial updates. This overhead can potentially erase the gains made by time slicing. To reduce the overhead, precalculated tables (the necessary start-up conditions) are kept.

Voice Activity Detection

The incoming signal is assumed to be either noise contaminated speech or noise alone. In order to accurately estimate the noise spectrum, the desired signal (speech) must be absent. Whenever noise alone is detected, a slowly averaged noise spectrum is updated. When speech is detected, the last updated noise spectrum is used to calculate the attenuation factors.

The voice activity detector is broadband and calculates two features:

1. Slowly decaying peak energy and
2. Minimum energy over 0.25 second intervals

In strongly modulated sections, indicating the presence of speech, the large energy excursions continually reset the peak energy to high values. The minimum energy level remains at the lowest excursions creating a wide gap between the two features. Conversely, in sections containing little modulation, the peak energy approaches the minimum energy.

To safeguard against false voicing detection, unvoiced must be declared for a number of consecutive frames – about one second. A counter accomplishes this operation starting at maximum count and decrementing until zero is reached. When zero is reached, an unvoiced detection is declared; otherwise, voicing is declared if a voiced frame occurs before this timeout operation. The counter is then reset back to maximum count.

This is a relatively simple feature to implement. Since power is at an extreme premium, it is necessary to keep only the best features.

Low Delay Noise Attenuation

This type of noise reduction is very effective in enhancing the quality of the signal. Since the channels are relatively wide, a separation between speech and noise is not possible; both speech and noise are attenuated by the same amount, therefore, this technique is best thought of as a remapping of speech and noise for the purposes of wearer comfort.

This noise attenuation algorithm uses two features to identify and attenuate noise in speech. Because this algorithm is aimed at hearing aid users, the artifacts and delay must be minimized while maximizing SNR and preceived speech quality.

Modulation

Modulation of speech is the first feature used to identify speech from noise. This feature was successfully used by Graupe and Causey in their noise attenuation algorithm [7]. Using modulation as a measure for SNR relies on the fact that the addition of stationary or nearly stationary noise to a speech

signal reduces the peak-to-peak modulation of the combined signal. **Figure 3** shows the fast RMS channel level for speech with no noise. **Figure 4** shows the fast RMS channel level for the same speech with additive white noise. The lower levels of the speech signal have been "filled-in", thereby reducing the difference between the maximum and minimum levels. As the SNR of the signal decreases, the difference between the maximum and minimum levels decreases. Using this method on bandlimited channels in the frequency domain, we arrive at **Figure 6**. It is a modulogram of clean speech with bandpassed additive white noise in channel 7 (2750Hz to 3250Hz). Each channel has a bandwidth of 500Hz, except for channel 1 and 16 which has a bandwidth of 250Hz.

Figure 5 shows a block diagram of the modulation detector. The modulation of a signal is measured by tracking the difference between the maximum and minimum values over time.

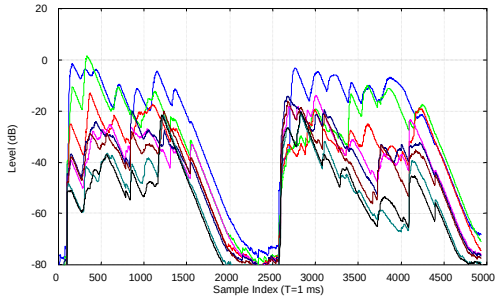

Figure 3: Fast RMS channel levels (τ=50 ms) for hint1a sentence ("The wife helped her husband. She's drinking from her own cup.") in quiet.

The maxima tracker is a peak detector with a first order exponential decay.

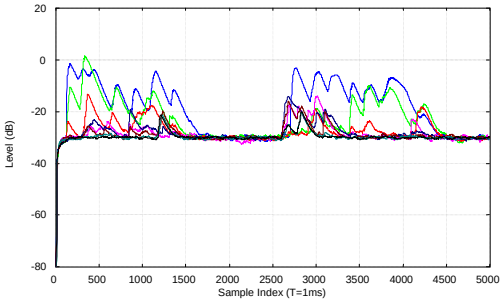

Figure 4: Fast RMS channel levels (τ=50 ms) for hint1a sentence ("The wife helped her husband. She's drinking from her own cup.") in noise.

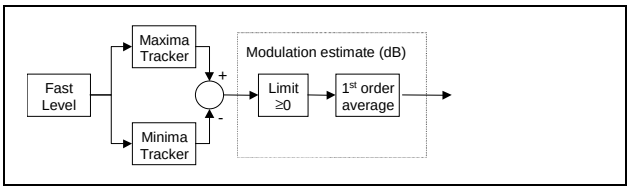

Figure 5: Modulation detector.

The minima tracker averages all local minimum values within 10ms windows in a 250ms time frame. An outlier rejector removes local minimum values that fall outside of

$$-2\sigma_{local\,\min ima} < \mu_{local\,\min ima} < \tfrac{1}{2}\sigma_{local\,\min ima},$$

where $\sigma_{local\,\min ima}$ is the standard deviation of the local minimum values within the 250ms time frame and $\mu_{local\,\min ima}$ is the mean local minimum value within the 250ms time frame, from the minimum tracker average.

This will track the noise level (instead of being biased by speech) while the outlier rejector ensures the estimate is not biased by transient signals [6].

Spectral Flux

The second feature we use is the rate of change of the power in each frequency channel. We will refer to this as the spectral flux [9]. Spectral flux is a relatively simple measure and can be thought of as a rudimentary voice detector.

The spectra flux is calculated as follows [8]. The power of each channel within a 10ms window is averaged. The difference is then calculated as $\frac{1}{12}(p[n+2]-8p[n+1]+8p[n-1]-p[n-2])$, where $p[n+2]$ is the 2nd 10ms window after $p[n]$. This difference is then averaged through a first order exponential formula.

Figure 6 shows the spectra flux for clean speech with additive bandpassed white noise from 2750Hz to 3250Hz.

Attenuation Rule

The modulation estimate and spectral flux are combined via a geometric mean and used as an input to an attenuation rule. Since decreasing modulation means a higher noise level, the rule applies an attenuation that is inversely proportional to modulation. The table is designed to minimize the attenuation of clean speech while presenting quickly increasing attenuation for signals with less than 10 dB modulation.

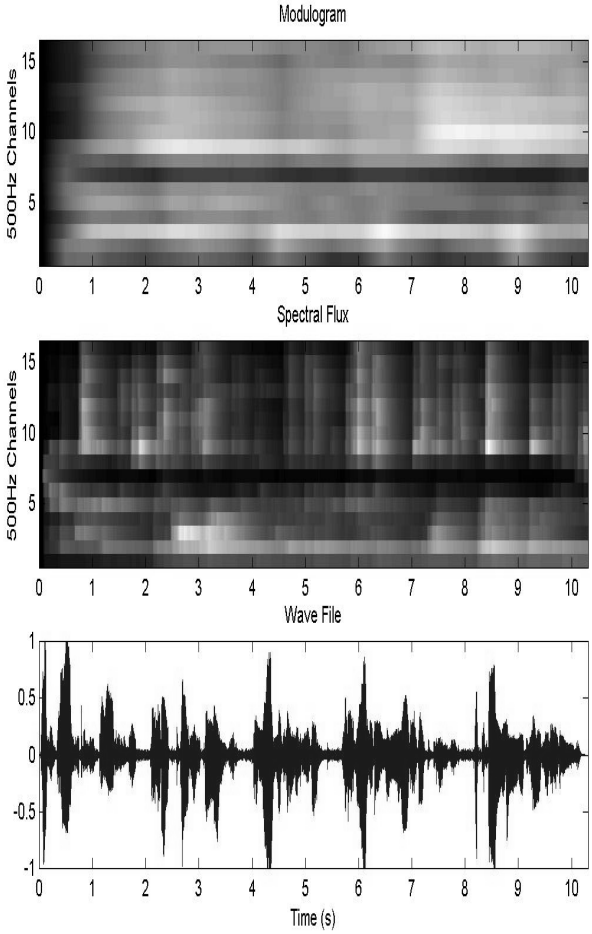

Figure 6: Modulogram and Spectral Flux of input signal

Results

The low delay noise attenuation algorithm was run on clean speech. There were no noticeable artifacts or degradations on the signal. The noise attenuation algorithm was then run on speech with additive bandpass stationary noise. The output waveform resulted in a substantial attenuation in the channel with the noise (>25dB attenuation) with little to no effect on other channels. The noise attenuation algorithm was then run on white noise, the white noise was attenuated substantially (>35dB) within 2 seconds.

The spectral subtraction routine was also run on clean speech. Again, there were no noticeable degradations perceivable. Wideband speech weighted noise was then added to the input speech. The resulting speech quality is quite good for SNR levels of 5dB or higher. At lower SNR levels, the algorithm has difficulty obtaining at accurate noise model resulting in a number of artifacts in the reconstructed speech. Further work is being pursued, borrowing the successful detection methods from the low delay noise attenuation algorithm, to build a voice activity detector capable of running at lower SNR levels.

References

[1] R. Brennan & T. Schneider, "Filterbank Structure and Method for Filtering and Separating an Information Signal into Different Bands, Particularly for Audio Signals in Hearing Aids", *PCT Patent Publication WO09847313A210*, October 22, 1998.

[2] R. Brennan & T. Schneider "A Flexible Filterbank Structure for Extensive Signal Manipulations in Digital Hearing Aids," *Proc. ISCAS-98*, Monterey, CA.

[3] R. E. Crochiere & L. R. Rabiner, "Multirate Digital Signal Processing", Prentice-Hall, 1983.

[4] T. Schneider & R. Brennan "A Multichannel Compression Scheme for a Digital Hearing Aid," *Proc. ICASSP-97*, Munich, Germany.

[5] P. P. Vaidyanathan, "Multirate Systems And Filter Banks", Prentice-Hall, 1993.

[6] E. Hänsler: "Acoustic Echo and Noise Control", *Proc. ICASP-99 Tutorial*, (1999).

[7] D. Graupe and G.D. Causey: "Method and Means for Adaptively Filtering Near-Stationary Noise from Any Information Bearing Signal." In US Patent 4,185,168 filed Jan. 4, 1978, expired 1980.

[8] P.V. O'Neil: Advanced Engineering Mathematics. Wadsworth Publishing Company, 1983

[9] M.J. Carey, E.S. Parris and H. Lloyd-Thomas: "A Comparison of Features for Speech, Music Discrimination." *Proc. ICASP 1999*, paper #1432