Speech Enhancement using an Adaptive Gain Equalizer

Nils Westerlund
Nils.Westerlund@bth.se

Mattias Dahl
Mattias.Dahl@bth.se

Ingvar Claesson
Ingvar.Claesson@bth.se

Department of Telecommunications and Signal Processing
Blekinge Institute of Technology
Ronneby, Sweden

June, 2003

Abstract

This paper presents a noise reduction method for speech communication where the input signal is divided into a number of subbands that are individually weighted in time domain according to the short time Signal-to-Noise Ratio estimate (SNR) in each subband at every time instant. Instead of focusing on suppression the noise, the method is focusing on speech enhancement.

The method has proven to be advantageous since it offers low complexity, low delay and low distortion. Also, there is no need for a Voice Activity Detector (VAD). The method is stand-alone and works regardless of speech coding schemes and other surrounding adaptive systems.

1 Introduction

Today, there are increasing demands for effective and secure inter-personal communication in various places and situations. Even in very noisy environments, it should be possible to communicate and human communication of today often occurs via some communication link. However, surrounding noise will degrade speech quality and intelligibility, forcing both the far-end and near-end user to strain both their hearing and their voices. In some situations, noise corrupted speech sent over a communication channel, can even be dangerous. An everyday application where low quality speech is harmless but nevertheless annoying, is the ordinary cellular phone, which is used in varying noisy environments such as in a moving car or in a crowd. Altogether, this urges for noise reduction methods.

A digital method for noise reduction in speech communication today, is spectral subtraction [1], [2]. The method is based on the Fourier Transform and is a non-linear, yet straightforward way of reducing unwanted broadband noise acoustically added to a signal. The noise bias is estimated during non-speech activity and then subtracted from the noisy speech spectra. Since the method estimates the noise bias during non-speech activity, a Voice Activity Detector (VAD) is required. One problem with spectral subtraction is musical tones but a number of enhancements of the original algorithm have been proposed during the years [3], [4], [5].

Multi microphone techniques such as microphone arrays, have also been investigated in order to spatially suppress a disturbance by adaptive beamforming [6].

This paper presents a general, stand-alone, time-domain method for speech enhancement, where the input signal is presented to a Single-Input-Single-

Output (SISO) system in which the signal is divided into a number of subbands. These subbands are then individually weighted in time domain according to the Signal-to-Noise Ratio (SNR) in each subband at every time instant. The method has proven to be robust, flexible and versatile and results in speech enhancement with low speech distortion without the requirement of a VAD.

2 Problem Formulation and Method

In a typical situation where a speech signal is distorted by noise , the noise, $w(n)$, is acoustically added to the speech, $s(n)$, by a microphone. The goal is to suppress the noise using some speech enhancement method resulting in an output signal, $y(n)$, with a higher SNR.

The basic idea behind the method described in this paper, is that a speech signal corrupted by bandlimited noise can be divided into a number of subbands and that each of these subbands can be individually and adaptively weighted according to an SNR estimate in that particular subband signal. A high subband SNR estimate indicates that the subband signal content is less corrupted by noise. Hence this subband should be boosted. A low subband SNR estimate indicates that the surrounding noise is dominant in the subband at hand. Hence no boosting of the subband speech should be performed.

To achieve this speech boosting effect, a slowly varying noise floor level estimate is calculated in each subband. A short term average is calculated simultaneously. Using the quotient of these quantities, a *gain function* can be achieved that weights the subband signal directly according to the subband signal SNR at that particular time instant. For example, if only noise is present in the signal, the noise floor level estimate and the short time average will be in the same order of magnitude, hence the quotient of these two measures will be unity. However, if speech is present, the short term average will increase but the noise floor level estimate will remain approximately unchanged. Hence, the quotient will become larger than unity, amplifying the signal in the subband at hand. It is also possible that a subband with a quite low SNR estimate, may contribute with some speech information during speech activity. Note that the method is focusing on speech enhancement rather than noise suppression.

Altogether, the advantages are manyfold when using this method, including increased speech quality. Also there is no need for a VAD. The method is flexible and versatile and in addition robust and stand-alone.

2.1 Mathematical Description

Suppose we have an acoustical noise denoted $w(n)$ and a speech signal denoted $s(n)$. The noise corrupted speech signal $x(n)$ can then be written as $x(n) = s(n) + w(n)$. By filtering the input signal $x(n)$ using a bank of K bandpass filters, $h_k(n)$, the signal is divided into K subbands, each denoted by $x_k(n)$ where k is the subband index. This filtering operation can be written in time domain as $x_k(n) = x(n) * h_k(n)$ where $*$ is the convolution operator. In the ideal case, the original signal can then be described as

$$x(n) = \sum_{k=0}^{K-1} x_k(n) = \sum_{k=0}^{K-1} s_k(n) + w_k(n) \qquad (1)$$

i.e. $x_k(n) = s_k(n) + w_k(n)$ where $s_k(n)$ is the speech part subband k and $w_k(n)$ is the noise part subband k. The output $y(n)$ is formed by

$$y(n) = \sum_{k=0}^{K-1} G_k(n) x_k(n) \qquad (2)$$

where $G_k(n)$ is a weighting function that amplifies the band k during speech activity. Since $G_k(n)$ introduces a gain to each subband, the function will be denoted *gain function* for the remainder of this paper.

Our desire is now to find a gain function that weights the input signal subbands using the ratio between $s_k(n)$ and $w_k(n)$, i.e. a short time SNR estimate. A block scheme illustrating the subband decomposition, weighting and final summation is shown in Fig. 1.

The gain function in each subband, is found by using a ratio of a short term exponential magnitude average, $A_{x,k}(n)$ based on $|x_k(n)|$, and an estimate of the noise floor level, $\underline{A}_{x,k}(n)$. The short term

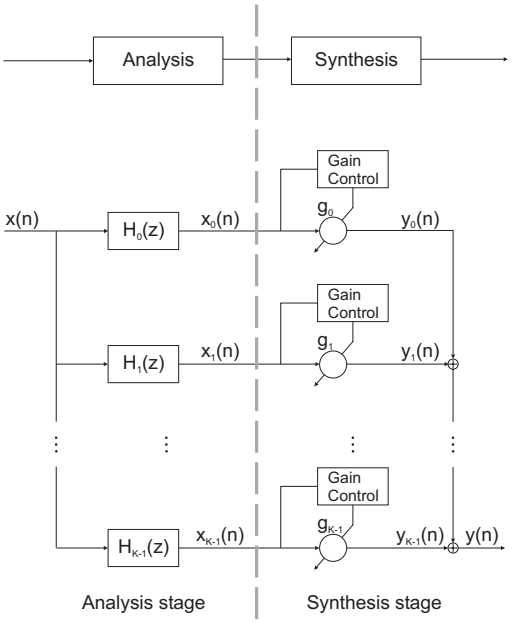

Figure 1: *Signal decomposition-weighting-assembly.*

average in subband k, $A_{x,k}(n)$, is calculated as

$$A_{x,k}(n) = \alpha_k A_{x,k}(n-1) + (1-\alpha_k)\,|x_k(n)| \quad (3)$$

where α_k is a small positive constant controlling how sensitive the algorithm should be to rapid changes in subband k input signal amplitude, i.e. a smoothing factor. A suitable value for α_k can be estimated using the following equation:

$$\alpha_k = \frac{1}{T_{s,k} F_s} \quad (4)$$

where F_s is the sampling frequency and $T_{s,k}$ is a time constant.

The slowly varying noise floor level estimate for each subband k, $\underline{A}_{x,k}(n)$, is calculated according to

$$\underline{A}_{x,k}(n) = \begin{cases} (1+\beta_k)\underline{A}_{x,k}(n-1) \text{ if case 1} \\ A_{x,k}(n) \text{ if case 2} \end{cases} \quad (5)$$

where 'case 1' corresponds to

$$\underline{A}_{x,k}(n-1) \leqslant A_{x,k}(n) \quad (6)$$

and 'case 2' corresponds to

$$\underline{A}_{x,k}(n-1) > A_{x,k}(n) \quad (7)$$

In (5), β_k is a small positive constant controlling how fast the noise floor level estimate in subband k will adapt to changes in the noise environment. Note that $A_{x,k}(n) \geqslant \underline{A}_{x,k}(n)$.

The variables $A_{x,k}(n)$ and $\underline{A}_{x,k}(n)$ are used to form the gain function $G_k(n)$ according to

$$G_k(n) = \left(\frac{A_{x,k}(n)}{\underline{A}_{x,k}(n)}\right)^{p_k}, \; p_k \geqslant 0, \; \underline{A}_{x,k}(n) > 0 \quad (8)$$

where p_k decides the amount of gain individually applied to each of the subband signals. The resulting speech enhanced output signal $y(n)$ is then calculated as

$$y(n) = \sum_{k=0}^{K-1} G_k(n) x_k(n) \quad (9)$$

Since the calculation of $G_k(n)$ involves a division, care must be taken to ensure that the quotient does not become excessively large due to a small $\underline{A}_{x,k}(n)$. For example, in a situation with a very high SNR, $G_k(n)$ will become very large if no limit is imposed on this function, resulting in an unacceptable high speech amplification. A limiter can be imposed on $G_k(n)$ as follows:

$$G_k(n) = \begin{cases} G_k(n) & \text{if} \quad G_k(n) \leqslant C_k \\ C_k & \text{if} \quad G_k(n) > C_k \end{cases} \quad (10)$$

where C_k is some positive constant.

An example of how the parameters $A_{x,k}(n)$, $\underline{A}_{x,k}(n)$ and $G_k(n)$ behave in a real situation is shown in Fig. 2.

3 Evaluation

A general mathematical description of the method was given in section 2.1 where a number of subband dependent parameters were introduced. In this section, where a practical evaluation is performed, some of these parameters are individually set to the same value for all subbands. Experiments have shown that using subband-dependent variables has no crucial effect on the final result. Nevertheless, subband-dependent variables can prove to be useful when tweaking the method.

All experimental evaluations has been performed on signals recorded on site at a sampling frequency of 8 kHz.

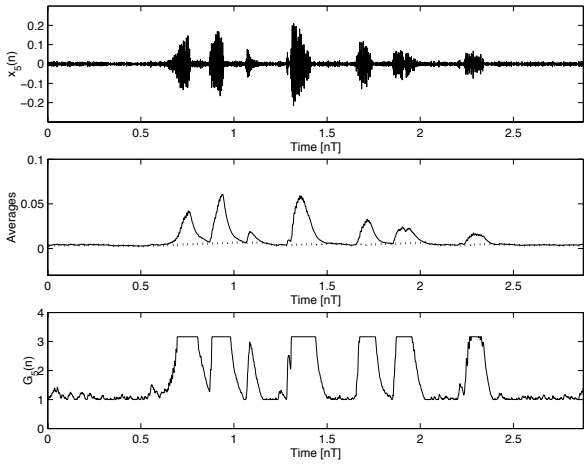

Figure 2: *(Upper plot) The speech signal $x_5(n)$ in subband 5 using a 12 band, 64 tap FIR-filter bank. (Mid plot) The solid line represents the short term average $A_{x,5}(n)$ with T_s=30 ms. The dashed line represents the corresponding noise floor level estimate $\underline{A}_{x,5}(n)$ with $\beta = 10^{-6}$. (Lower plot) The resulting gain function $G_5(n)$ limited to a maximum amplification of 10 dB. In all plots $T = 1/Fs$*

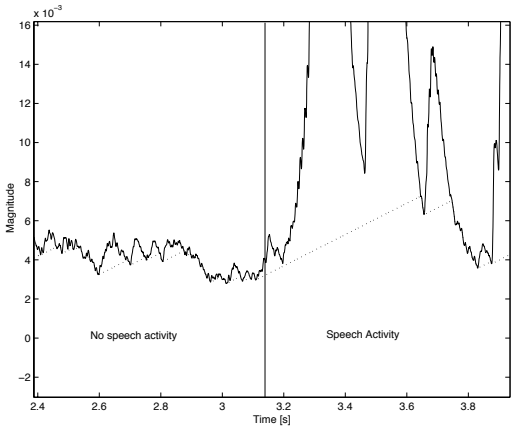

Figure 3: *Detailed behavior of short term average (solid line) and noise floor level estimate (dotted line) during both speech and non-speech activity.*

3.1 The Filter Bank

In section 2.1, a bank of bandpass filters was used to divide the input signal into k subbands. Intentionally, this was a quite general description.

The FIR filters used in this paper are designed with the window method using Hamming windows. This method results in causal, symmetric impulse responses with linear phase.

3.2 The Short Term Average

In order to effectively calculate a short term average magnitude of the input signal, a recursive structure was used as described in section 2.1, (3) and (4). A very small value of T_s results in unnatural sounding speech and background noise. If T_s is chosen to be much larger, the short term average reacts to slowly to incoming speech resulting in poor speech enhancement performance. An extremely large T_s will result in a function approximately equal to the noise floor level estimate described in section 2.1, (5). Hence, $G_k(n)$ will be equal to unity in each subband and no speech amplification will be carried out.

3.3 The Noise Floor Level Estimate

The noise floor level estimate described in section 2.1, (5), is designed to adaptively track the background noise level. The value of $\underline{A}_{x,k}(n)$ is dependent on the short term average, $A_{x,k}(n)$. The positive constant β controls how fast the noise floor level estimate will adapt to changes in the noise environment. A very small value of β results in a noise floor level estimate approximately equal to the short term average while a very large β results in slow convergence and poor noise level tracking capabilities in non-stationary environments.

3.4 The Gain Function

As described in section 2.1, (8), the gain function $G_k(n)$ is the ratio of the short term average and the noise floor level estimate. Since we are dealing with a ratio, care must be taken to avoid singularities. The remedy to this problem is to use an upper limit imposed on the gain function $G_k(n)$ as described in section 2.1, (10).

An exponential weighting of $G_k(n)$ is performed by raising the ratio $A_{x,k}(n)/\underline{A}_{x,k}(n)$ to a power of p, as in section 2.1, (8).

An illustration of the detailed behavior of the short term average and the noise floor level estimate both during speech and non-speech activity is shown in Fig. 3.

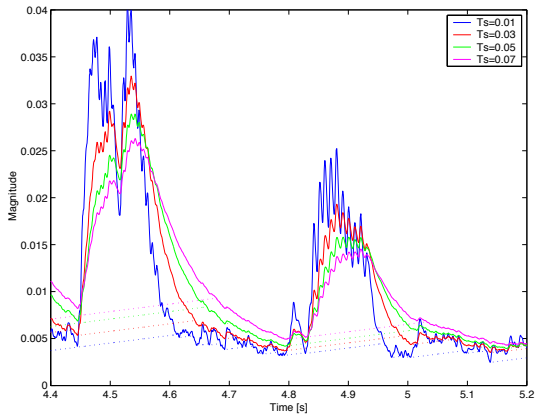

Figure 4: *The short term average (solid lines) and noise floor level estimate (dotted lines) calculated with four different values of T_s.*

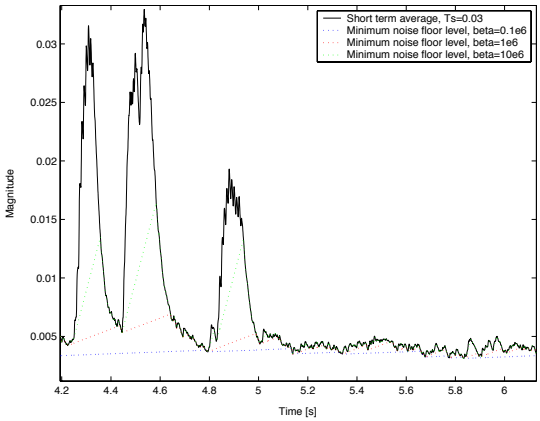

Figure 6: *Effects on the noise floor level estimate when altering the parameter β.*

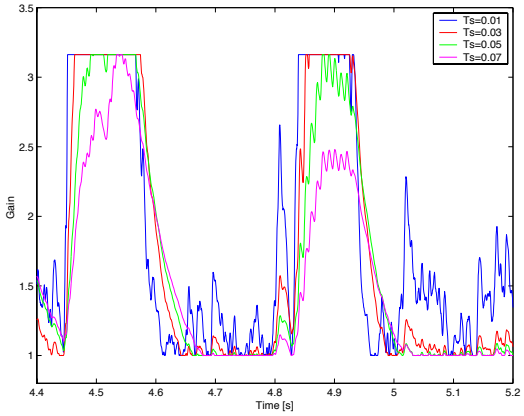

Figure 5: *The gain function for four different values of the time constant T_s.*

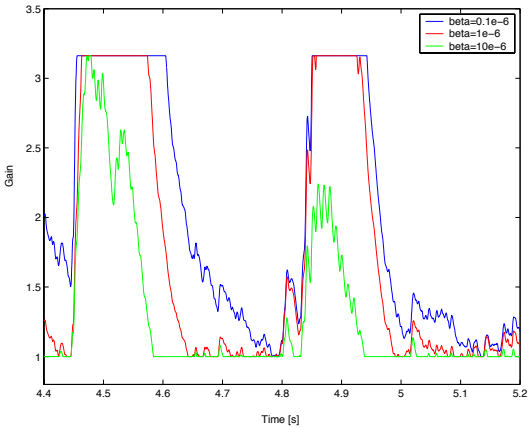

Figure 7: *Effects on the gain function when altering the parameter β.*

4 Results

The choice of the constants α and β is crucial to the speech enhancement performance and the resulting speech quality.

As mentioned before, a small α results in unnatural sounding speech with remaining artifacts. A very large α results in a short term average that reacts too slowly to incoming signal amplitude variations. Hence, the speech attacks will be cropped and, in addition, the speech amplification in subbands with a small amount of noise will be limited.

These conclusions are illustrated in Fig. 4 and Fig. 5.

The parameter β also has a fundamental effect on the final result. A small β results in a noise floor level estimate that reacts rapidly to changes in the noise environment. The disadvantage is that the minimum noise floor also reacts to incoming speech. A larger β results in a more stable minimum noise floor estimate but also in slower convergence and poor noise level tracking capabilities in non-stationary environments. The different effects on the noise floor level estimate when altering the parameter β are illustrated in Fig. 6 and Fig. 7.

4.1 Speech Enhancement Performance

As mentioned earlier, the method does not focus on *noise suppression* but rather on *speech enhancement*. It is important to notice the fact that we are actually not trying to improve the SNR by removing noise. Instead, the SNR is improved by amplifying the speech on a subband basis.

Experimental results shows that the number of subbands is not crucial for most types of noise. However, a more complex noise environment, e.g. narrow band noise or even periodic disturbances, may be more efficiently suppressed using a larger number of subbands.

Furthermore, the time constants that controls the short term average should be kept in the range of speech pseudo-stationarity time, i.e. about 20–30 ms [2]. The noise floor level estimate controlling parameter, β, can be varied around 10^{-6} depending on the desired effects. The upper bounding of the gain function $G(n)$ affects the resulting speech distortion and should be kept within 5–20 dB. A larger amplification of the signal may result in a piercing sounding output speech.

5 Conclusions

A speech enhancement algorithm of today must be flexible yet robust, easily implemented yet computationally efficient and, in addition, it must be versatile and applicable to many different noise situations. The method described in this paper fulfills all of these criterions: Thanks to the straightforward basic underlying idea, the method is both flexible, robust and easily implemented. The filter bank structure used in this paper was an FIR filter bank but principally, any filter bank could be employed to obtain desired features and characteristics. In addition, experiments has shown that the method is easily configured and adapted to different noise situations with a minimum of tweaking.

Finally, the fact that the need for a VAD is eliminated, makes the method highly robust and a strong competitor (or complement) to existing noise reduction methods.

References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-27, pp. 113–120, April 1979.

[2] J. R. Deller Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete time processing of speech signals*, Macmillan Publishing Company, 1993.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-32, pp. 1109–1121, December 1984.

[4] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 2, pp. 345–349, 1994.

[5] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 8, 2001.

[6] M. Dahl and I. Claesson, "Acoustic noise and echo canceling with microphone array," *IEEE Trans. On Vehicular Technology*, vol. 48, no. 5, pp. 1518–1526, September 1999.