A SPECTRAL SUBTRACTION RULE FOR REAL-TIME DSP IMPLEMENTATION OF NOISE REDUCTION IN SPEECH SIGNALS

Matteo Romanin
Dept. of Information Engineering
University of Padova, Padova, Italy
matteo.romanin@dei.unipd.it

Enrico Marchetto
Dept. of Information Engineering
University of Padova, Padova, Italy
enrico.marchetto@dei.unipd.it

Federico Avanzini
Dept. of Information Engineering
University of Padova, Padova, Italy
federico.avanzini@dei.unipd.it

ABSTRACT

Spectral subtraction is a method for restoration of the spectrum magnitude for signals observed in additive noise, through subtraction of an estimate of the average noise spectrum from the noisy signal spectrum. In this paper we show that, starting from the known minimum mean-square error (MMSE) suppression rules of Ephraim and Malah and under the same modeling assumptions, a simpler suppression filtering rule can be found. Moreover, we demonstrate its performances and compare its computational costs with respect to the reference rule of Ephraim and Malah. This result permits a real time implementation of the exposed theory with an efficient algorithm on the DSP TMS320 C6713B.

1. INTRODUCTION

Noise reduction systems are used for many applications where various sources of ambient broadband noise degrade the quality of acquired audio signals. A great deal of work has been spent for automatic speech and/or speaker recognition, and the literature about the state of the art is more and more focused about noisy environments [1]. Moreover, since long time it is known the negative impact of noise on the recognition over the telephone channel [2, 3].

Many short-time spectral attenuation techniques may be found in the literature, in which a time-varying (zero phase) filter, or suppression rule, is applied to the short-time Fourier transform of a corrupted signal with the goal of improving the acquired SNR.

The minimum mean-square error (MMSE) suppression rules due to Ephraim and Malah [4] appears to be extremely effective. Unfortunately, direct implementation of this rule requires a great amount of computations, which make it not suitable for real time implementations. In particular the estimation of the suppression filter implies computation of Bessel functions.

In this paper we start from the rule of Ephraim-Malah in order to find a simpler function to weight the spectrum of noisy speech signal. The proposed filtering rules is characterized by a very low discrepancy with respect to the Ephraim-Malah solution and by a dramatically lower computational load. In addition we compare our rule to a similar simplified rule (Minimum Mean-Square Error Spectral Power Estimator), originally presented in [5].

Finally a real-time DSP implementation of the proposed noise-suppression approach is discussed. The implementation has been obtained with the floating-point digital signal processor TMS320 C6713B.

2. EPHRAIM-MALAH SUPPRESSION RULE

In this section we briefly summarize the suppression rule presented in [4]. We assume that $y(n)$, the noise corrupted discrete input signal, is composed of the clean speech signal $x(n)$ plus the uncorrelated additive noise signal $d(n)$:

$$y(n) = x(n) + d(n). \tag{1}$$

The signal is processed on a short-time basis (frame-by-frame) in the frequency domain:

$$Y_k(i) = X_k(i) + D_k(i), \tag{2}$$

where Y_k, X_k, D_k are the discrete Fourier transforms of y, x, d; the index $k \in [0, N)$ represents the k-th bin of the spectrum (with N the order of the DFT) and the index $i \in \mathbb{Z}$ the i-th time frame.

Noise reduction is achieved by the application of a suppression rule, the nonnegative real valued gain $H_k(i)$, to each bin k of the observed signal spectrum $Y_k(i)$ in order to obtain an estimate $\hat{X}_k(i)$ of the clean speech $X_k(i)$:

$$\hat{X}_k = H_k(i)\, Y_k(i). \tag{3}$$

2.1. Additive Gaussian Model

In a Gaussian Model assumption the spectral components of Y_k and D_k are modeled as independent, zero-mean, complex Gaussian random variables (for simplicity index i is omitted):

$$Y_k = |Y_k|\, e^{\jmath\angle Y_k} \in \mathcal{N}\left(0, \sigma_y^2\right) \tag{4}$$

$$D_k = |D_k|\, e^{\jmath\angle D_k} \in \mathcal{N}\left(0, \sigma_d^2\right) \tag{5}$$

with variances σ_y^2 and σ_d^2. In such hypothesis the probability density for the absolute value of the complex variable Y_k is modeled by the Rayleigh function:

$$p_{\{|Y_k|\}}(a_k) = \begin{cases} \frac{1}{\sigma_y^2}\ a_k\ \exp\left(\frac{-a_k^2}{2\sigma_y^2}\right) & x \geq 0, \\ 0 & x < 0. \end{cases} \tag{6}$$

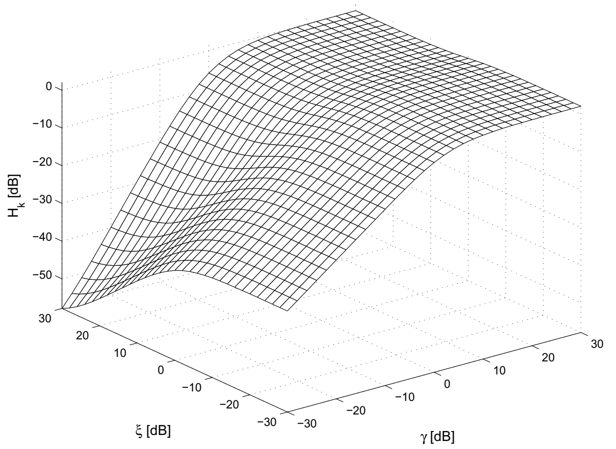

Figure 1: *Ephraim-Malah suppression rule H_k (in dB) for different values of a priori and a posteriori SNR (ξ and γ resp.).*

Similarly for the spectrum D_k of the noise signal we have:

$$p_{\{|D_k|\}}(a_k) = \begin{cases} \frac{1}{\sigma_d^2}\, a_k\, \exp\left(\frac{-a_k^2}{2\sigma_d^2}\right) & x \geq 0, \\ 0 & x < 0. \end{cases} \quad (7)$$

2.2. MMSE Suppression Rule

The Ephraim-Malah MMSE log estimator is a short-time spectral amplitude estimator H_k that minimizes the mean-square error of the estimated logarithms of the spectra $\hat{X}_k$. It takes the following form:

$$H_k = \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\left\{\left[(1+\nu_k)\, I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] e^{-\frac{\nu_k}{2}}\right\}, \quad (8)$$

where $I_n(\cdot)$ is the modified Bessel function (MBF) of order n and where ν_k is defined as follow:

$$\nu_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k. \quad (9)$$

The function ξ_k is the *a priori* signal-to-noise ratio and the function γ_k is the *a posteriori* signal-to-noise ratio. They are given by:

$$\xi_k \triangleq \frac{|X_k|^2}{|D_k|^2}, \qquad \gamma_k \triangleq \frac{|Y_k|^2}{|D_k|^2}. \quad (10)$$

Figure 1 shows the reference rule as a function of the parameters ξ and γ expressed in dB.

2.3. *A priori* SNR, *a posteriori* SNR

The computation of the *a priori* SNR and the *a posteriori* SNR requires the knowledge of the clean speech spectrum X_k, which is not available. An estimation of the *a posteriori* SNR can be obtained as:

$$\gamma_k(i) \triangleq \frac{|Y_k(i)|^2}{\hat{D}_k(i)}, \quad (11)$$

under the assumption that the noise $d(n)$ is stationary and that an estimation $\hat{D}_k$ of its power spectrum may be computed during portions of the input signal $y(n)$ where no speech is present:

$$\hat{D}_k(i) \triangleq (1-\beta)\ |Y_k(i)|^2 + \beta\, \hat{D}_k(i-1)\Big|_{x(n)=0}. \quad (12)$$

The $\beta \in [0,1]$ parameter in (12) controls the update speed of the recursion from frame $(i-1)$-th to frame i-th.

An estimation of the *a priori* SNR can be obtained with a decision-directed approach [5]:

$$\xi_k(i) \triangleq (1-\alpha)\hat{\gamma}_k(i) + \alpha\frac{|H_k(i-1)Y_k(i-1)|^2}{\hat{D}_k(i)}, \quad (13)$$

where $\alpha \in [0,1]$ parameter rules the update speed of the recursion at the same manner as in (12) and where the function $\hat{\gamma}_k$ is simply defined as:

$$\hat{\gamma}_k = \begin{cases} \gamma_k & \gamma_k \geq 0, \\ 0 & \gamma_k < 0. \end{cases} \quad (14)$$

2.4. Computational cost

In this section we compute the number of operations used by the implementation of the exact Ephraim-Malah rule on the DSP. The largest computational load is due to the MBFs of order 0 and 1, $I_n(\nu/2)$. For integer values of n, the MBFs are defined as:

$$I_n(\nu) = \left(\frac{\nu}{2}\right)^n \sum_{m=0}^{+\infty} \frac{\left(\frac{\nu}{2}\right)^{2m}}{m!(m+n)!}. \quad (15)$$

We need to find the index M where the summation can be stopped, in such a way that the truncation error is smaller than machine precision. The TMS320 C6713B DSP used in this work is featured with a 32-bit floating-point single precision CPU, where $1.17549 \cdot 10^{-38}$ is the smallest representable normalized number. The denominator in Eq. (15) leads the summatory terms to zero; $M = 34$ may be used as a good truncation point, as we have $1/(2(M!)) = 1.6936 \cdot 10^{-39}$.

To achieve a better estimate for the truncation point M we draw in Fig. 2 the absolute error E_M between a double precision "exact" implementation, computed using the MBFs available in Matlab (double precision), and our implementation, truncated and with single precision. There are three sources of error which affect only the single precision algorithm:

- the single precision itself;
- the truncation error for the formula (15);
- the saturation of ν[1].

Fig. 2 uses $M = 1000$ to keep low the truncation error and highlight the others; we see the little numerical noise and, more noticeable, the saturation error in the high-SNRs area. Numerical and saturation errors can not be eliminated, so we proceed to the optimization of M without taking care of these. Summing over ξ and γ the squared error for various M values we are able to find the optimal M used for truncation: $M = 68$. This is the lowest one able to keep the whole error exactly the same as $M = 1000$ or more (i.e. with $M = 68$ we keep only the irremovable errors).

[1] This is due to the term $\exp(-\nu_k/2)$ in (8); when ν exceeds 177 this term becomes too small to be represented in single precision, thus causing the whole formula to diverge.

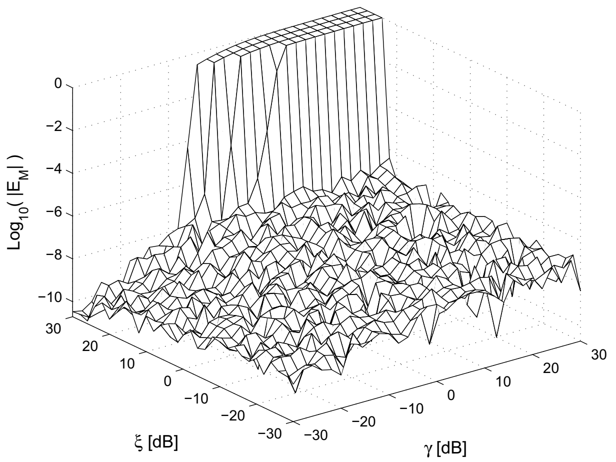

Figure 2: *Logarithm of the absolute error E_M for* (8) *computed using single and double precision; see text for details.*

Having chosen a suitable M, we are now able to count the floating-point operations used by our algorithm; in the following we suppose the cost of any sum and product operation equal to 1. The Ephraim-Malah rule (8) is computed for every bin of the FFT, which are twice as much as the cardinality of the samples in the time domain because of the 50% overlap of the audio frames.

For each time sample the formula (8), in our implementation for single precision and with 50% frames overlap, uses: two exponentials, two square roots and $2\cdot(14+11M) = 1524$ sum/product operations.

3. PROPOSED SUPPRESSION RULE

The spectral amplitude estimator given by (8) requires the computation of exponential and Bessel functions. This great amount of operations is not suitable for real-time implementations, as it is hardly sustainable also by highest performance DSPs.

In the past a number of simplified rules has been proposed; in particular in [5] there are three alternative rules, each one supported by a different theoretical view. Our simplified formula is geared toward a fast real-time DSP implementation with a reasonably little performance loss over (8).

3.1. An approximated suppression rule

We propose a suppression rule that approximates the Ephraim-Malah rule as the root of a second-order polynomial. The MBFs $I_n(\cdot)$ have the following asymptotic forms for non-negative integer n:

$$\begin{aligned} I_n(x) &\sim \frac{1}{\sqrt{2\pi x}} e^x & x &\gg \left|n^2 - 1/4\right|, \\ I_n(x) &\sim \frac{1}{n!}\left(\frac{x}{2}\right)^{2n} & 0 &< x \ll \sqrt{n+1}, \end{aligned} \tag{16}$$

which show that the the expression between braces in Eq. (8) is asymptotic to $\sqrt{\nu_k}$ for $\nu_k \to +\infty$, while it tends to 1 for $\nu_k \to 0$.

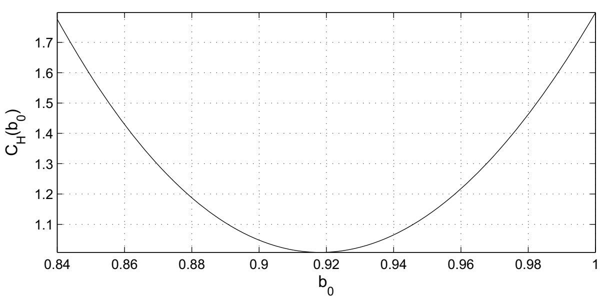

Figure 3: *The global minimum of the cost function $C_H(b_0)$ found for $b_0 = 0.9182$.*

In light of this observation we introduce the following approximated suppression rule:

$$\hat{H}_k = \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\sqrt{b_0 + b_1\nu_k}, \tag{17}$$

where the coefficients $b_{0,1}$ have to be determined in order to minimize the error of $\hat{H}_k$ with respect to H_k. Equivalently, we seek $b_{0,1}$ to provide the best fit of the line $b_0 + b_1\nu_k$ to the function:

$$\left[(1+\nu_k)\, I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right]^2 e^{-\nu_k}. \tag{18}$$

Using the first equation in (16) for $n = 0, 1$, the coefficient b_1 can be determined by observing that the function (18) has the following asymptotic behavior:

$$\left[(1+\nu_k)\, I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right]^2 e^{-\nu_k} \sim \frac{4}{\pi}\nu_k, \tag{19}$$

for $\nu_k \to +\infty$. In practice this approximation holds for $\nu_k \gg 3/4$ (see Eq. (16)). By comparison with Eq. (17), one can see that in order for $\hat{H}_k$ to fit H_k the value:

$$b_1 = \frac{4}{\pi} \tag{20}$$

has to be chosen. *Viceversa*, for small values of ν_k, expanding the exponential function in (18) according to its Maclaurin series up to the second order, and expanding the MBFs according to the definition (15), yelds the approximation:

$$\left[(1+\nu_k)\, I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right]^2 e^{-\nu_k} \sim 1 + \nu_k + \frac{\nu_k^2}{2}, \tag{21}$$

for $0 < \nu_k \ll 1$. It follows that in this limit:

$$\hat{H}_k \sim \frac{\sqrt{\pi\nu_k}}{2\gamma_k}\sqrt{\left[1 - \left(\frac{4}{\pi} - 1\right)\nu_k + \frac{\nu_k^2}{2}\right] + \frac{4}{\pi}\nu_k}. \tag{22}$$

The expression in brackets in Eq. (22) suggests that the interval where to seek the optimal value of b_0 is a left neighborhood of the value 1.

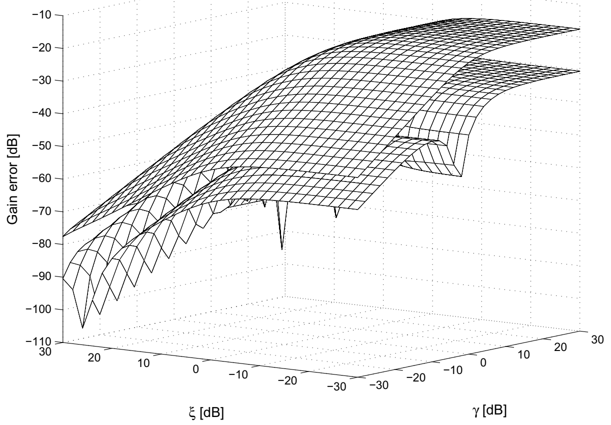

Figure 4: *Gain errors of $\tilde{H}_k$ (upper surface) and $\hat{H}_k$ (lower surface) with respect to the reference H_k.*

3.2. Optimal coefficient b_0

A key point in the approximation $\hat{H}_k$ in (17) is the choice of the optimal coefficient b_0. We find the optimal b_0 by minimizing the square error between $\hat{H}_k$ and H_k. That is, the optimal b_0 minimizes the cost function:

$$C(b_0) \triangleq \int_{\xi}\int_{\gamma} |E_k|^2 \, d\xi d\gamma, \tag{23}$$

where the error function $E_k = H_k - \hat{H}_k$ depends parametrically on b_0. In Fig. 1 the two independent variables ξ and γ both vary in the range $[-30, 30]$ dB, similarly to [5]. We choose to compute the cost function $C(b_0)$ in these ranges of values. Leaving implicit the conversion to linear values, and referring to (9) to obtain ν, we rewrite the cost function as:

$$C(b_0) = \sum_{\xi=-30}^{30}\sum_{\gamma=-30}^{30}\left[H_k - \hat{H}_k(b_0)\right]^2, \tag{24}$$

with the only independent variable b_0. The optimization is thus straightforward; we use a (linear) grid approach, refining the grid in the vicinity of the global minimum indicated by Eq. (22) and related statements. Fig. 3 shows the cost function in the range $[0.84, 1.0]$; the value

$$b_0 = 0.9182 \tag{25}$$

is found to provide the absolute minimum for $C(b_0)$. Recalling Eq. (20) and rearranging terms yields the approximated suppression rule used in the DSP implementation:

$$\hat{H}_k = \frac{1}{\gamma_k}\sqrt{\frac{\pi}{4}\left(b_0\nu_k + b_1\nu_k^2\right)} = \frac{1}{\gamma_k}\sqrt{0.7212\nu_k + \nu_k^2}. \tag{26}$$

3.3. Gain error and computational costs

The proposed rule $\hat{H}_k$ in Eq. (26) has been compared quantitatively to the original Ephraim-Malah rule H_k by studying the dB error ($20\log_{10}|E_k|$) over the whole ranges for ξ and γ.

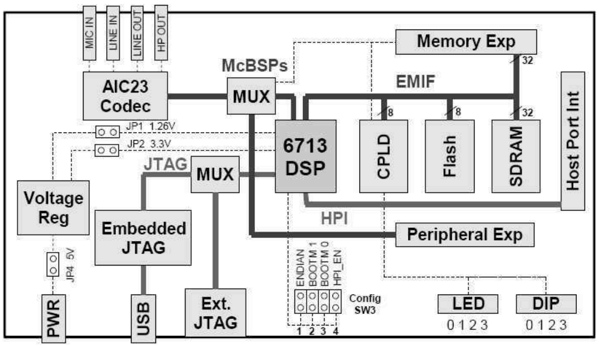

Figure 5: *Block diagram TMS320 C6713B DSK board.*

Additionally we have compared our approximated rule with the MMSESP estimator proposed by Wolfe and Godsill in [5]:

$$\tilde{H}_k = \sqrt{\frac{\xi_k}{1+\xi_k}\left(\frac{1+\nu_k}{\gamma_k}\right)} = \frac{1}{\gamma_k}\sqrt{\nu_k + \nu_k^2}. \tag{27}$$

The reason is that this latter estimator is functionally very similar to our Eq. (26) (with different b_0 and b_1 values), although it has been found based on a statistical rather than algebraic approach.

Figure 4 shows the gain differences of both our rule and $\tilde{H}_k$ with respect to the reference H_k. From this figure one can note that, despite the approximations introduced and the limitations due to single precision, the absolute maximum error provided by $\hat{H}_k$ is well below -30 dB over the whole SNR ranges and below -40 dB in the most interesting areas ($\xi > 0$ dB). Moreover, comparison of the two error surfaces shows that the gain error of $\tilde{H}_k$ is higher throughout the considered range, and reaches a maximum at about -18 dB.

The computational cost of the proposed approximation (26) is, for each time-domain sample, the following: two square roots and 10 operations (5 operations counted twice because of the 50% overlap). The MMSESP estimator (27) uses two square roots and $4 \cdot 2 = 8$ operations; thus our approximation carryes only a little extra computational effort. By contrast you can compare these numbers with those in the end of section 2.4.

3.4. DSP Implementation

A real-time implementation of the proposed suppression rule was obtained with the board DSK TMS320 C6713B, equipped with the floating-point digital signal processor TMS320 C6713B (Fig. 5). Operating at 225 MHz, this DSP delivers up to 1350 million floating-point operations per second (MFLOPS), and 1800 million instructions per second (MIPS). The CPU fetches advanced very-long instruction words (256 bits wide) to supply up to eight 32-bit instructions to the eight functional units during every clock cycle.

Referring to Fig. 6, the DSP receives the analog audio signals through an on-board TLV320 AIC23 codec with 90 dB SNR Multibit Sigma-Delta ADC (A-weighted at $F_s = 48$ kHz). The input signal is segmented in blocks of $2N = 512$ samples with an overlap of 256 samples (50%) based on Hanning window. The frame-rate period is equal to $T_f = 5.33$ ms; consequently the maximum DSP load capacity is fixed to 9.6 million instructions per frame.

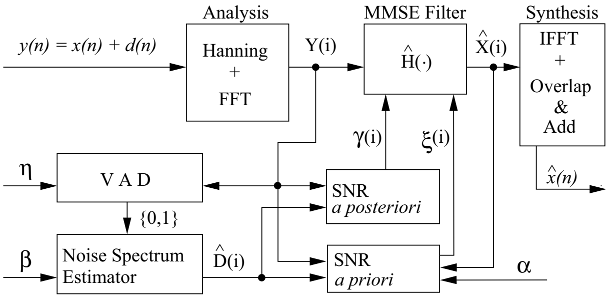

Figure 6: *Schematic of the DSP implementation.*

Spectral analysis is implemented efficiently by means of a $N = 256$ complex points radix-2 FFT. The original $2N$-point real sequence is packed as a N-point complex sequence, on which N-point complex FFT is applied. The resulting N-point complex output is unpacked into another $N + 1$ point complex sequence, which corresponds to spectral bins 0 to N of the $2N$-point real input sequence.

After filtering by means of our suppression rule, the inverse FFT and the Overlap-Addition method are used to obtain segments of processed speech, which are passed to the D/A converter. A voice activity detector (VAD) is required to identify those frames of the input signal in which only noise is present; the noise spectrum is updated only in these frames using Eq. (12). The VAD has to accurately identify frames of silence in order to avoid erroneous updates of the noise spectrum including parts of the speech signal. The detector used in this implementation is based on a statistical model-based voice activity detection approach; it computes the likelihood ratio of speech being present or absent in the input frame as described in [6]. The parameter η in Fig. 6 determines the threshold of speech level detection.

Table 1 reports results about the computational load in million instructions per frame, for both the original rule H_k and the proposed rule $\hat{H}_k$. The benchmark profiler integrated in DSP Code Composer Studio v3.3 was used in order to obtain these estimates. The number of cycles refers to a single input frame. As discussed in Sec. 2.4, the CPU load of the original rule is computed in single-precision using $M = 68$, by englobing the exponential function $e^{-\nu/2}$ in (8) into the Bessel series, and expanding the term $(\nu/2)^{2m+n}$ in (15) as a product of single factors to preserve the numerical accuracy.

	Max Cycles [mill. instr.]	Avrg. Cycles [mill. instr.]
Ephraim-Malah Rule H_k	17.697	17.586
Proposed Rule $\hat{H}_k$	0.221	0.153

Table 1: *DSP computational loads per frame ($2N = 512$).*

The computational loads reported in Table 1 show that direct implementation of the Ephraim-Malah suppression rule is far from providing a real-time de-noising algorithm on the DSP TMS320 C6713B. On the other hand the approximated rule can straightforward be implemented on DSP without further optimization, as it requires a small fraction of the available computational power.

4. CONCLUSIONS

We have proposed a new suppression rule for noise reduction applications looking at the well-known Ephraim-Malah rule as a reference, but aiming toward real-time DSP implementations. In the first part of the paper we have determined the minimum number of iterations needed to obtain the precise computation of the MBFs for the Ephraim-Malah rule. This analysis has provided a quantitative estimate of the computational cost of the Ephraim-Malah rule. Moreover the analysis has shown that double precision calculations are needed for accurate implementation.

The proposed rule is obtained starting from asymptotic considerations on the Ephraim-Malah rule, based on which a simplified equation with two coefficients is found and a range for possible values for these coefficients is determined. We have then applied a straightforward numerical optimization procedure to fine-tune the parameters. The results discussed in Sec. 3.3 show that, despite the apparently crude approximations, the proposed rule exhibits negligible gain errors.

We have then described the computational cost for our proposed rule, with references to a real-world DSP implementation. A real-time set up has been developed in order to perform direct comparisons between computational loads of the original and approximated rules. This comparison shows that the approximated rule requires a small fraction of the computational power of the DSP. On the contrary the DSP is not able to support the original Ephraim-Malah rule in real-time.

We have done some tests using real-world audio samples and pointing out the good properties of our real-time implementation. Our future activities will try to apply some objective measures, like those of the PESQ standards, which should show the improved intelligibility of the filtered speech. Combining noise suppression and speech enhancement may also give interesting results.

5. REFERENCES

[1] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero, "Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 1061–1070, July 2008.

[2] Pedro J. Moreno and Richard M. Stern, "Sources of degradation of speech recognition in the telephone network," in *ICASSP*, April 1994, vol. 1, pp. 109–112.

[3] Douglas A. Reynolds, M. A. Zissman, Thomas F. Quatieri, G. C. O'Leary, and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *ICASSP*, May 1995, vol. 1, pp. 329–332.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

[5] Patrick J. Wolfe and Simon J. Godsill, "Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement," *J. App. Sig. Proc.*, vol. 2003, no. 10, pp. 1043–1051, 2003.

[6] M. Bhatnagar Y. Hu and P. Loizou, "A cross-correlation technique for enhancing speech corrupted with correlated noise," in *ICASSP*, May 2001, vol. 1, pp. 673–676.