SPEECH ENHANCEMENT USING A MINIMUM LEAST SQUARE AMPLITUDE ESTIMATOR

Christophe Beaugeant[1] and Pascal Scalart[2]

[1] SIEMENS AG, ICM, Grillparzerstrasse 10-18, 81675 Munich, GERMANY
[2] France Telecom R&D, 2 av. Pierre Marzin, 22307 Lannion cedex, FRANCE
{e-mail: christophe.beaugeant@mch.siemens.de, pascal.scalart@rd.francetelecom.fr}

ABSTRACT

In this contribution, we address the general problem of noise reduction for speech signal. We propose a new filter designed in the frequency domain by taking into account actual analysis frame but also previous frames. We compared our solution with the classical Wiener filter designed by the analysis of a unique actual frame. The proposed approach provides also new highlights on the empirical Ephraim and Malah *a priori* SNR estimator which is commonly used in practice.

1. INTRODUCTION

Noise is part of our daily reality but if noise can give an pleasant sensation (feeling of an atmosphere), too much noise or aggressive noise involve tiredness and misunderstanding of speech. As a result, in many applications such as telecommunication or hearing aids, noise reduction is mandatory.

Most of the actual solutions are based on a well-known family of speech enhancement algorithms: *short-time spectral attenuation algorithms* where the analysis is performed in the frequency domain. They attempt to estimate the short-time spectral magnitude of the speech $\hat{S}(m, f_k)$ by applying an attenuation $G(m, f_k)$ to each short-time Fourier transform coefficients of the noisy speech $Y(m, f_k)$ at frame *m*:

$$\hat{S}(m, f_k) = G(m, f_k)\, Y(m, f_k)$$

where f_k represents the *k*th spectral component.

Various noise reduction systems have been proposed depending on the chosen suppression rule. In order to provide a common theoretical basis for relating these algorithms, it has been found useful to analyze these noise reduction filters for a given frame of data of length $T \approx 20\text{ms}$ where quasi-stationarity of the speech is ensured. Conversely, most of these noise reduction systems use in practice estimators that are computed from the knowledge of previous frames (for example magnitude averaging). The algorithm presented in this article has the advantage of taking into account the previous frames directly during the filter's design, involving less empirical estimations. In Section 2 we derived the Least Square (LS) amplitude estimator and we discuss about the asymptotic behavior of the filter. We also discuss the explicit connection of the proposed LS approach to the so-called Ephraim and Malah *a priori* SNR estimator proposed in [1]. This interpretation will be discussed in Section 3. Finally, Section 4 presents comparative experimental results of this new algorithm.

2. LEAST SQUARE AMPLITUDE ESTIMATOR

In this Section, we analyze the spectral attenuation problem through the minimization of the LS criterion. It has the advantage to focus our analysis on the succession of input frames whereas most noise suppression rules are based on the analysis of a single frame of length *T*. To analyze the LS approach, we assume that all processes are stationary, and we introduce the following assumptions:

(A1) The Fourier expansion coefficients of each process is statistically independent

(A2) The observation signal is considered as the sum of a useful signal $S(m, f_k)$ and a noise signal $N(m, f_k)$

(A3) The spectral components of the useful signal and the noise are uncorrelated.

2.1. Optimal LS filter

Let us defined a cost function as the weighting of the squared errors computed over successive frames *l*:

$$J_m(e(f_k)) = \sum_{l=0}^{m} \lambda^{m-l} e^2(l, f_k) \qquad (1)$$

with the error on the frame *l* defined as :

$$e(l, f_k) = S(l, f_k) - \hat{S}(l, f_k)$$

Note the analogy with the LS algorithm applied in the time domain, with the difference that we deal here with a succession of frames in the frequency domain.

With the following notations for vectors:

$$\mathbf{E}(m, f_k) = [\lambda^{\frac{m}{2}} e(0, f_k); \lambda^{\frac{m-1}{2}} e(1, f_k); ...; e(m, f_k)]^T$$

$$\mathbf{S}(m, f_k) = [\lambda^{\frac{m}{2}} S(0, f_k); \lambda^{\frac{m-1}{2}} S(1, f_k); ...; S(m, f_k)]^T$$

$$\hat{\mathbf{S}}(m, f_k) = [\lambda^{\frac{m}{2}} \hat{S}(0, f_k); \lambda^{\frac{m-1}{2}} \hat{S}(1, f_k); ...; \hat{S}(m, f_k)]^T$$

$$\mathbf{Y}(m, f_k) = [\lambda^{\frac{m}{2}} Y(0, f_k); \lambda^{\frac{m-1}{2}} Y(1, f_k); ...; Y(m, f_k)]^T \qquad (2)$$

the error function becomes :

$$J_m(e(f_k)) = \mathbf{E}^H(m, f_k)\mathbf{E}(m, f_k) \qquad (3)$$

Under assumption (A1), and assuming that the phase of the noisy speech is not processed, based on the assumption that phase distortion is not perceived by the human ear, we can write :

$$\hat{\mathbf{S}}(m, f_k) = \mathbf{Y}(m, f_k)\, G_{\mathrm{LS}}(m, f_k)$$

It is mandatory to note that in this last expression, for a given frequency f_k, $G_{\mathrm{LS}}(m, f_k)$ is a scalar function of the frame index *m*, whereas $\mathbf{E}(m, f_k)$, $\mathbf{S}(m, f_k)$. $\hat{\mathbf{S}}(m, f_k)$ and $\mathbf{Y}(m, f_k)$ are vectors of dimension *m*.

Considering (2) and (3) the cost function $J_m(e(f_k))$ can be expressed thanks to the variables $\mathbf{S}(m, f_k)$, $G_{\mathrm{LS}}(m, f_k)$, and $\mathbf{Y}(m, f_k)$ as followed :

$$J_m(e(f_k)) = (\mathbf{S}(m, f_k) - \mathbf{Y}(m, f_k)G_{\mathrm{LS}}(m, f_k))^H$$
$$(\mathbf{S}(m, f_k) - \mathbf{Y}(m, f_k)\, G_{\mathrm{LS}}(m, f_k))$$

The resolution of the equation $\partial J_m / \partial G_{\mathrm{LS}} = 0$ gives us the filter $G_{\mathrm{LS}}(m, f_k)$ minimizing the error function J_m. This leads to the following equation :

$$\mathbf{Y}^H(m, f_k)\mathbf{Y}(m, f_k)G_{\mathrm{LS}}(m, f_k) = \mathbf{Y}^H(m, f_k)\mathbf{S}(m, f_k)$$

As a result when $\sum_{l=0}^{m} \lambda^{m-l} Y^2(l, f_k) \neq 0$ (which is the case for non silent periods), we can write :

$$G_{LS}(m, f_k) = \frac{\sum_{l=0}^{m} \lambda^{m-l} Y^*(l, f_k) S(l, f_k)}{\sum_{l=0}^{m} \lambda^{m-l} Y^2(l, f_k)} = \frac{Num(m, f_k)}{Den(m, f_k)}$$

Using assumption (A2) in the numerator and denominator of this equation leads to :

$$Num(m, f_k) = \sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k) + \sum_{l=0}^{m} \lambda^{m-l} S(l, f_k) N^*(l, f_k)$$

$$Den(m, f_k) = \sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k) + \sum_{l=0}^{m} \lambda^{m-l} N^2(l, f_k) + 2\,\mathrm{Re}\left(\sum_{l=0}^{m} \lambda^{m-l} S(l, f_k) N^*(l, f_k)\right)$$

Moreover, assuming ergodicity, we can notice that :

$$\lim_{m \to \infty} \sum_{l=0}^{m} \lambda^{m-l} S(l, f_k) N^*(l, f_k) = \frac{1}{1-\lambda} \gamma_{sn}(m, f_k)$$

where $\gamma_{sn}(m, f_k)$ is the cross-power spectrum between the noise and the useful signal. As a result, when $m >> 1$ and by exploiting the hypothesis (A3) that the useful signal and the noise are uncorrelated ($\gamma_{sn}(f_k) = 0$), the following approximation of the filter $G_{\mathrm{LS}}(m, f_k)$ is obtained :

$$G_{\mathrm{LS}}(m, f_k) \approx \frac{\sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k)}{\sum_{l=0}^{m} \lambda^{m-l} S^2(l, f_k) + \sum_{l=0}^{m} \lambda^{m-l} N^2(l, f_k)} \qquad (4)$$

2.2. Asymptotic behavior

One can note the similarity between $G_{\mathrm{LS}}(m, f_k)$ in (4) and the Wiener filtering [2] expressed as:

$$G_{\mathrm{Wiener}}(m, f_k) = \frac{\gamma_{ss}(m, f_k)}{\gamma_{ss}(m, f_k) + \gamma_{nn}(m, f_k)} = \frac{SNR_{\mathrm{prio}}(m, f_k)}{1 + SNR_{\mathrm{prio}}(m, f_k)} \qquad (5)$$

where the *a priori* Signal to Noise Ratio (SNR) is define as:

$$SNR_{\mathrm{prio}}(m, f_k) = \frac{\gamma_{ss}(m, f_k)}{\gamma_{nn}(m, f_k)}$$

Indeed, the filter expressed in (4) can be seen as a Wiener filter where the estimations of the power spectral densities $\gamma_{ss}(f_k)$ and $\gamma_{nn}(f_k)$ are defined by the following expression (ergodicity hypothesis), with $m >> 1$:

$$\hat{\gamma}_{uu}(m, f_k) = (1-\lambda) \sum_{l=0}^{m} \lambda^{m-l} \left| U(l, f_k) \right|^2, \; U \in \{S, N\} \qquad (6)$$

If we also define the *a priori* Signal to Noise Ratio estimator as :

$$\overset{\wedge}{SNR}_{\mathrm{prio}}(m, f_k) = \frac{\hat{\gamma}_{ss}(m, f_k)}{\hat{\gamma}_{nn}(m, f_k)} \qquad (7)$$

for $m >> 1$, (4) leads to similar formula as in (5):

$$G_{\mathrm{LS}}(m, f_k) = \frac{\hat{\gamma}_{\mathrm{ss}}(m, f_k)}{\hat{\gamma}_{\mathrm{ss}}(m, f_k) + \hat{\gamma}_{\mathrm{nn}}(m, f_k)} = \frac{\overset{\wedge}{SNR}_{\mathrm{prio}}(m, f_k)}{1 + \overset{\wedge}{SNR}_{\mathrm{prio}}(m, f_k)}$$

In this last expression, it is remarkable that the estimations of psd are derived from (4), *i.e.* during the initial design phase of the filter, whereas they do not appear directly in most of noise suppression rules since their analysis is based on a single frame. Most of the time, the problem of estimating the parameters which are involved in a given suppression rule is often empirical, and is addressed in a second step after the derivation of the theoretical filter expression. In the approach presented in this paper, they are directly defined through the mathematical design of the weighting rule. In the next section, we lay out the stress on this property by demonstrating that the *a priori* SNR estimator proposed in [1] can be partially rediscovered thanks to the previous LS analysis.

3. A PRIORI SNR ESTIMATOR

The derivation of the *a priori* SNR estimator is based on the definition (7) which can be expressed as:

$$\overset{\wedge}{SNR}_{\text{prio}}(m,f_k)=\frac{S^2(m,f_k)}{\sum_{l=0}^{m}\lambda^{m-l}N^2(l,f_k)}+\lambda\frac{\sum_{l=0}^{m-1}\lambda^{m-1-l}S^2(l,f_k)}{\sum_{l=0}^{m}\lambda^{m-l}N^2(l,f_k)} \qquad (8)$$

Using in the above equation the following equality $\gamma_{nn}(m,f_k)=(1-\lambda)\sum_{l=0}^{m}\lambda^{m-l}N^2(l,f_k)$, we have also :

$$\overset{\wedge}{SNR}_{\text{prio}}(m,f_k)=(1-\lambda)\frac{S^2(m,f_k)}{\gamma_{nn}(m,f_k)}+\lambda\frac{\sum_{l=0}^{m-1}\lambda^{m-1-l}S^2(l,f_k)}{\sum_{l=0}^{m}\lambda^{m-l}N^2(l,f_k)} \qquad (9)$$

Inserting (4) in the right-hand term of this relation, and assuming that $m>>1$, it follows that

$$\frac{\sum_{l=0}^{m-1}\lambda^{m-1-l}S^2(l,f_k)}{\sum_{l=0}^{m}\lambda^{m-1-l}N^2(l,f_k)}=\frac{G_{\text{LS}}(m-1,f_k)}{1-G_{\text{LS}}(m-1,f_k)} \qquad (10)$$

$$=\frac{G_{\text{LS}}(m-1,f_k)Y(m-1,f_k)}{\left[1-G_{\text{LS}}(m-1,f_k)\right]Y(m-1,f_k)}$$

Moreover, making the following approximation :

$$Y(m,f_k)\approx S(m,f_k)+\sqrt{\gamma_{nn}(m,f_k)}$$

which is similar to the maximum likelihood estimate of the speech spectrum in the method of the spectral amplitude subtraction, we can write:

$$\begin{aligned}\left[1-G_{\text{LS}}(m-1,f_k)\right]Y(m-1,f_k)&\approx\left(S(m-1,f_k)+\sqrt{\gamma_{nn}(m-1,f_k)}\ \right)\\&\quad-G_{\text{LS}}(m-1,f_k)Y(m-1,f_k)\\&=\left(S(m-1,f_k)-\hat{S}(m-1,f_k)\right)+\sqrt{\gamma_{nn}(m-1,f_k)}\\&\approx\sqrt{\gamma_{nn}(m-1,f_k)}\end{aligned} \qquad (11)$$

Combining (11) and (10), and substituting (10) in (9), and using the following definition for the *a posteriori* SNR

$$SNR_{\text{post}}(m,f_k)=\frac{S^2(m,f_k)}{\gamma_{nn}(m,f_k)}$$

it can be verified that the *a priori* SNR can be estimated through a recursive approach which is given by

$$\begin{aligned}\hat{SNR}_{\text{prio}}(m,f_k)&=(1-\lambda)\ SNR_{\text{post}}(m,f_k)\\&\quad+\lambda\ \frac{G_{\text{LS}}(m-1,f_k)\ Y(m-1,f_k)}{\sqrt{\gamma_{nn}(m-1,f_k)}}\end{aligned} \qquad (12)$$

Noting that $\hat{S}(m-1,f_k)=G_{\text{LS}}(m-1,f_k)\ Y(m-1,f_k)$, we see that relation (12) uses the amplitude estimator of the $(n-1)^{\text{th}}$ frame instead of the amplitude itself in the *n*th frame. Thus, the proposed estimator (12) corresponds to a "decision-directed" approach, since $\hat{SNR}_{\text{prio}}(m,f_k)$ is updated on the basis of a previous amplitude estimate. It is interesting to note that Ephraim and Malah find the following corresponding expression for the *a priori* SNR estimator [1]:

$$\begin{aligned}SNR_{\text{prio}}(m,f_k)&=(1-\beta)\ SNR_{\text{post}}(m,f_k)\\&\quad+\beta\ \frac{{G_p}^2(m-1,f_k)Y^2(m-1,f_k)}{\gamma_{nn}(m,f_k)}\end{aligned} \qquad (13)$$

which was obtained from an empirical weighted averaging between two potential estimates of the *a priori* SNR, and was found to be very useful when it is combined with the MMSE amplitude estimator (see [1], [3]). Comparing (12) and (13), we see that the difference consists only on taking into account amplitudes instead of squared amplitudes in the second term of both expressions. The influence of this difference is discussed in the next section. The interest of such a similarity between relations (12) and (13) lies in the fact that the *a priori* SNR estimator proposed in [1] can be interpreted as resulting from a minimization of a least squared criterion.

Such an interpretation allows a better understanding of the Ephraim and Malah short-time amplitude (EMSA) estimator. Optimal values of the parameter β are usually found by simulations only. To provide an efficient noise reduction and an enhanced speech with *colorless* residual noise, values of β sufficiently close to one ($\beta\approx 0{,}96-0.98$) are usually required ([1], [3]). From our LS interpretation, such values correspond to the introduction in (1) of a forgetting factor λ near unity. Thus, the cost function $J_m(.)$ is mainly influenced by the previous successive short-time frames and, as a result, time variations of the noise reduction filter in (4) are highly smoothed. These remarks were already given in [3] where the behavior of the EMSA estimator is investigated. However, the least square approach highlights the understanding of the mechanisms that counter the musical noise by considering a cost function $J_m(.)$ with exponential windowing of the frame-by-frame error sequence $\{e(l,f_k)\}$.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

To compare the estimators introduced in the previous section, we propose in this section to analyze the behavior of the Wiener filter $G_{Wiener}(m,f_k)$ presented in (5) as a function of the $SNR_{prio}(m,f_k)$ (either thanks to the EMSA estimator (13) or to the one derived from the LS analysis (12)). The figures Fig-1 shows the variations of $G_{\text{Wiener}}(m,f_k)$ as a function of the *a posteriori* Signal to Noise Ratio $SNR_{post}(f_k)$ and of the following SNR

$SNR_{prev}(m-1, f_k) = \frac{G_{\text{Wiener}}{}^2(m-1, f_k)\, Y^2(m-1, f_k)}{\gamma_{nn}(m-1, f_k)}$. This last parameter can be considered as a Signal to Noise Ratio estimated on the previous frame $(m-1)$.

These graphs show that the behavior of the filter $G_{\text{Wiener}}(m, f_k)$ is few dependant on the estimator when $SNR_{\text{post}}(m, f_k)$ or $SNR_{\text{prev}}(m, f_k)$ are greater than -5 dB.

Conversely, when both SNRs are low, the attenuation provided by the EMSA estimator is greater than the one derived from the least square analysis. That means that when the noise is preponderant in the noisy signal $Y(m, f_k)$, the least square analysis will give a smaller attenuation than the empirical analysis of EMSA estimator.

Nevertheless, except this last case, the behavior of G_{Wiener} is not so affected by the square factor difference previously noted in Section 3 between the two estimators. This practically justified low difference between the two estimators provides a true justification of the use of the LS analysis as an explanation of (13). It also justifies the remarks and comments made in the previous section.

5. SUMMARY

In this article, a new method for noise reduction is proposed. It is based on the minimization of the LS amplitude criterion computed in the frequency domain on the successive short-time frames of analysis. The proposed approach has the advantage of taking into account the previous frames directly during the filter's design, involving less empirical estimations than common noise suppression rules. Thanks to few assumptions, we have shown that the LS approach highlights the understanding of the mechanisms that counter the musical noise phenomenon and also enlightens the so-called Ephraim and Malah *a priori* SNR estimator. Finally, experimental results given for the Wiener suppression rule, show that the SNR estimator derived from the proposed LS analysis provides similar behavior than the Ephraim and Malah *a priori* SNR estimator. As a result by using the LS filter described in this article, the same filtering properties obtained by combining Wiener filter and Ephraim and Malah estimator are expected, namely a good trade-off between the level of musical noise in the enhanced speech and the distortion brought on the useful speech.

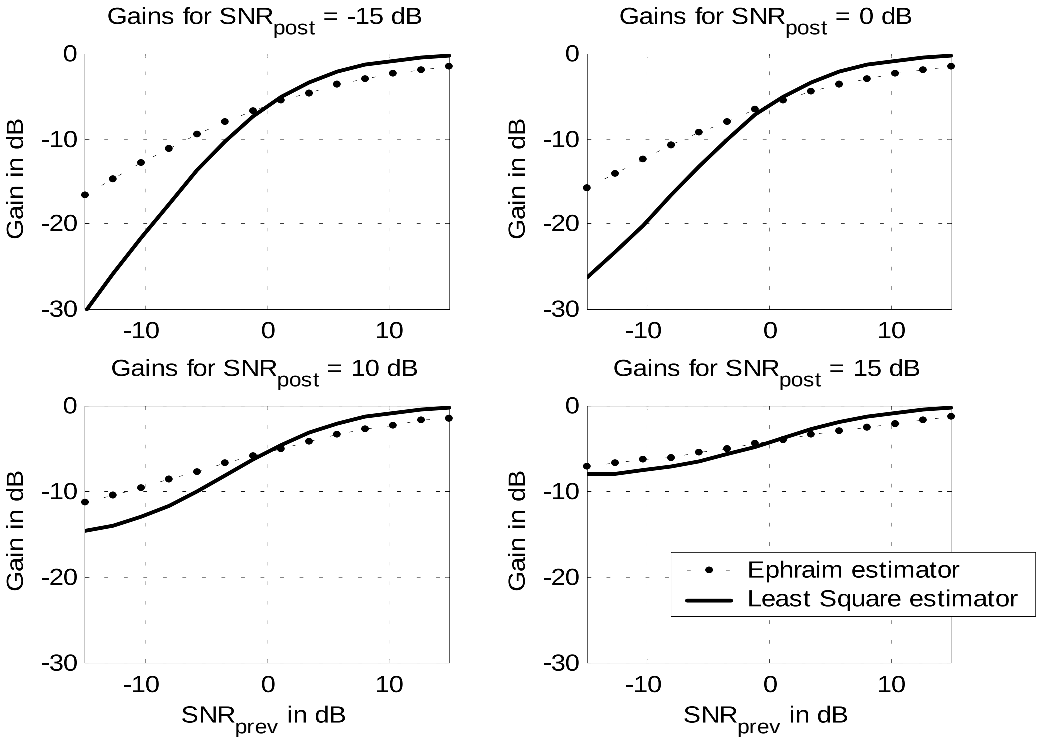
6. REFERENCES

[1] Y. Ephraim, D. Malah, "Speech Enhancement using Optimal Non-linear Spectral Amplitude Subtraction," in Proc. *ICASSP'83*, pp. 1118-1121, 1983.

[2] S.V. Vaseghi. *Advanced Signal Processing and Digital Reduction*, Chapter 9, Wiley Teubner Communication, Queen's University Belfast, UK, 1996.

[3] O. Cappé, "Elimination of Musical Tone with the Ephraim and Malah Suppressor," in *IEEE Trans. ASSP*, No.2, pp. 345-349, April 1994.

Fig.1– Parametric gain curves describing Wiener gain functions behavior using estimators defined by (12) and (13).


Gains for SNR$_{post}$ = -15 dB
Gains for SNR$_{post}$ = 0 dB
Gains for SNR$_{post}$ = 10 dB
Gains for SNR$_{post}$ = 15 dB
Gain in dB
SNR$_{prev}$ in dB
Ephraim estimator
Least Square estimator