On Speech Enhancement Algorithms Based on MMSE Estimation

Pascal SCALART[1], Jozue VIEIRA FILHO[2,3], José GERALDO CHIQUITO[3]

[1]FRANCE TELECOM - CNET LAA/TSS/CMC
Technopole ANTICIPA, 2 Avenue Pierre Marzin, 22307 Lannion Cedex, FRANCE

[2]Universidade Estadual Paulista DEE/FEIS/UNESP, Av. Brasil Centro 56, Ilha solteira- SP, BRAZIL

[3]Universidade Estadual de Campinas (DECOM/FEE/UNICAMP), SP, BRAZIL

E-mail : scalart @lannion.cnet.fr

ABSTRACT

This paper addresses the problem of single microphone frequency domain MMSE noise reduction technique for speech enhancement in noisy environments. We first analysed asymptotic performance of the MMSE estimate and compared these results with the Wiener filter. Practical implementation of the MMSE filter is then presented. Comparisons between optimal and practical behaviour of the MMSE filter demonstrate that an effective improvement in the noise reduction process can be gained if greater attention is given to the these estimators.

1 INTRODUCTION

To date, many single microphone noise reduction techniques are based on the assumption that it is mainly the spectral magnitude rather than the phase that is important for speech intelligibility and quality. In such systems, the noisy speech is first windowed and then transformed in the frequency domain. The enhanced spectral magnitude is evaluated on each frequency according to a short-time suppression rule. The enhanced speech signal is then recovered by inverse transforming this spectral magnitude estimation combined with the phase of the noisy speech signal.

Many approaches have been proposed for the evaluation of the short-time suppression factor. This factor is adjusted individually on each frequency as a function of the local signal to noise estimation. Such methods include power spectral subtraction [1], Wiener filtering [2], soft-decision estimation [3] and Minimum Mean Square Error (or MMSE) estimation [4]. An important question is : What is the best choice for this short-time suppression rule in order to provide the best results in terms of speech quality and intelligibility, nature of residual noise and amount of noise reduction ?

Analysis of literature shows that MMSE estimate has recently received much attention [6, 8] by many researchers for speech enhancement in the context of mobile hands-free radio communications. However, asymptotic properties of this estimate (i.e. with optimal choice of the internal parameters of the estimation process) have never been explored. In this paper, the asymptotic behaviour of the MMSE short-time spectral amplitude estimator is evaluated and compared to the Wiener one in order to provide information on the improvement in the enhanced speech quality that can be gained with such suppression rules. In the second part, we recall the different estimates of the filter parameters that are usually implemented for real-time operations. We then compare this practical implementation performance with optimal evolution of the internal parameters of the MMSE suppression rule. Our results, based on objective measures and informal subjective tests, show that significant improvement can be gained if greater attention is given to the filter parameters estimation.

2 MMSE SHORT-TIME SPECTRAL ESTIMATE

Let *s(t)* and *b(t)* denote the speech and the additive noise processes, respectively. The observed signal *x(t)* is given by $x(t) = s(t) + b(t)$. Let $S_k = A_k\, e^{j\,\alpha_k}$, B_k, $X_k = R_k\, e^{j\,\nu_k}$, denote the *k*th spectral component of the signal *s(t)*, the noise *b(t)* and the noisy observations *x(t)* in the analysis interval [0,T] where quasi-stationarity of speech signal is guaranteed over the time period T.

The MMSE short-time spectral amplitude estimate proposed by Ephraim and Malah [4, 10] makes it possible to derive a solution to the speech enhancement problem by determining a more fundamental theoretical analysis than conventional systems like wiener or power spectrum subtraction suppression rule.

Under the assumed MMSE statistical model, the phase of the clean speech signal in each spectral bin is uniformly distributed on the interval $[0,2\pi]$ and the spectral amplitude has a Rayleigh probability density function. The MMSE amplitude estimate $\hat{A}_k$ of the speech is thus evaluated from X_k by a non-linear gain function defined by $G(f_k) \triangleq \hat{A}_k / X_k$ which is expressed as the product of the standard gain by a term which contributes to the "soft-decision" aspect of the estimate as given by :

$$G(f_k) = \frac{\Lambda(f_k)}{1+\Lambda(f_k)} \frac{\sqrt{\pi}}{2} \sqrt{\frac{1}{SNR_{post}} \frac{SNR_{prio}}{1+SNR_{prio}}}\, F\left[SNR_{post} \left(\frac{SNR_{prio}}{1+SNR_{prio}} \right) \right]$$

$$\text{where } F[x] = \exp(-x/2)\left[(1+x) I_o\left(\frac{x}{2}\right) + x I_1\left(\frac{x}{2}\right) \right] \qquad (1)$$

with $I_0(.)$ and $I_1(.)$ denote the modified Bessel functions of zero and first order.

The parameter $\Lambda(f_k)$ is the generalised likelihood ratio taking into account the uncertainty of speech presence in the noisy observations defined by :

$$\Lambda(f_k) \;=\; \mu_k \frac{p\left(X_k / H_k^1\right)}{p\left(X_k / H_k^0\right)} \qquad (2)$$

with $\mu_k \triangleq (1-q_k)/q_k$, where q_k is the probability of signal absence in the *k*th spectral component, and *p(.)* denotes a probability density function. H_k^0 and H_k^1 denote the two hypotheses of signal absence and presence, respectively, in the *k*th spectral component.

In the previous definition of the MMSE amplitude estimate the local *a posteriori* and *a priori* SNRs are given by :

$$SNR_{post}(f_k) \triangleq \frac{|X_k|^2}{E\left\{|B_k|^2\right\}} \qquad SNR_{prio}(f_k) \triangleq \frac{E\left\{|S_k|^2\right\}}{E\left\{|B_k|^2\right\}} \qquad (3)$$

3 ANALYSIS OF THE MMSE SUPPRESSION RULE

In this section, the asymptotic performance of the MMSE filter are analysed and compared to the Wiener one which is given by :

$$H_{Wiener}(p, f_k) = \frac{SNR(p, f_k)}{1+SNR(p, f_k)} \qquad (4)$$

In order to evaluate the optimal behaviour of these filters, we have made the implicit assumption that the time and frequency domain evolution of the internal parameters of spectral gain functions are known. The principle is shown in Figure 1.

The clean speech and the noisy signal are used separately for the evaluation of the filters parameters : the SNR for the Wiener filter and the a posteriori and a priori SNR for the MMSE suppression rule. Furthermore, in order to provide information on the quality of the enhanced speech signal we have also computed objective measures between the clean speech signal and the output of the filter.

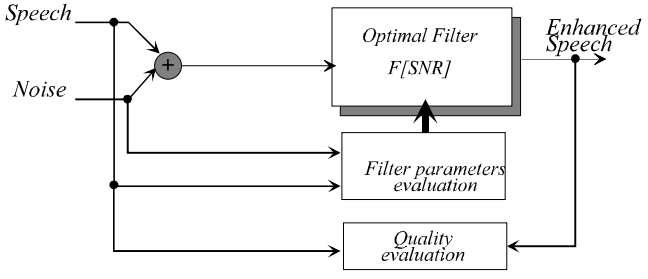

Figure 1. Optimal process evaluation method.

Experiments have been made with speech corrupted by background noise recorded in vehicule. Typical results are shown on figure 2 where we have represented the noisy (a) speech signal and the output of the filters for the MMSE (b) and Wiener (c) suppression rules. The mean value of the input signal to noise ratio is 5 dB. During non-speech activity periods, we can see that the Wiener filter achieves a better noise reduction than the MMSE suppression rule. The residual noise power of the Wiener filter is lower than with the MMSE approach. This observation is also true in the low frequency components when speech is present.

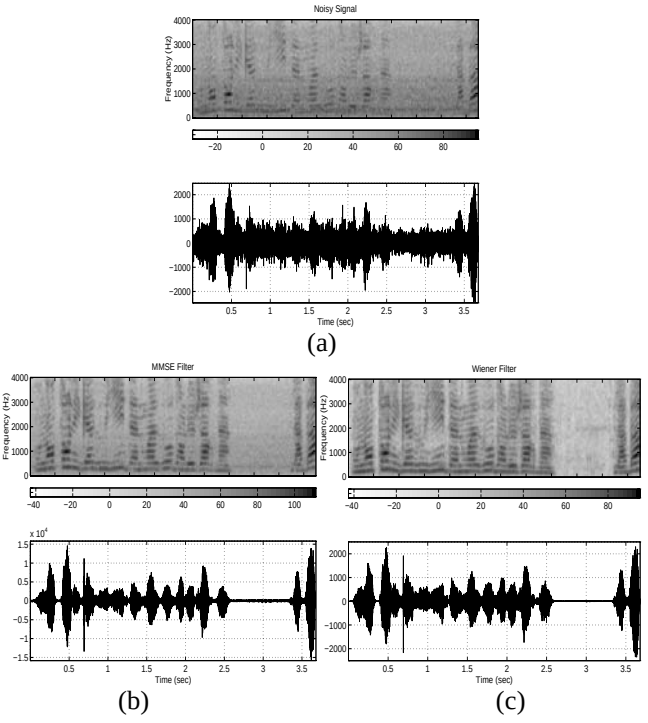

Figure 2. (a) Noisy speech signal (mean SNR of 5 dB) and output of the optimal MMSE (b) and Wiener (c) filters.

In order to provide information on speech distortions, we have represented in Figure 3 the mean value of cepstral and basilar distances between the clean speech and the enhanced speech signals at the output of the two filters.

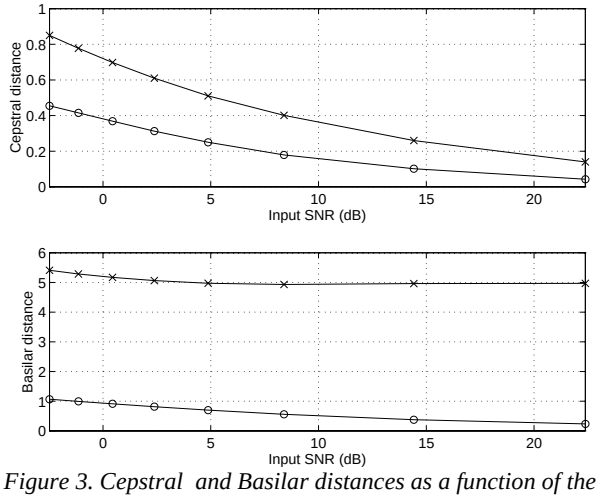

Figure 3. Cepstral and Basilar distances as a function of the input SNR for optimal MMSE (-x-) and Wiener (-o-) filters.

These measures are represented for different values of the mean input SNR. The basilar distance is expressed in dB and corresponds to a perceptual objective measure with is evaluated through a modelisation of the human ear in order to provide the excitation pattern on the basilar membrane. The cepstral distance is derived from the LPC coefficients during speech activity periods.

Analysis of cepstral and basilar curves indicate that the MMSE suppression rule gives always an enhanced speech signal which presents more distortions than the Wiener one.

One should also note that this conclusion holds even for input signal to noise greater than 20 dB. Further experiments on larger sample of speech confirm these results.

They can be partly explained through the analysis of real speech magnitude distribution. If we look at the assumptions of the statistical MMSE model [4], we can see that the Fourier expansion coefficients of speech are modelled as statistically independent gaussian random variables.

However, this assumption is not really true. Figure 4 shows the cumulative distribution of real speech spectral magnitude evaluated at different frequencies on large sample of clean speech (male and female speakers). We have also represented the cumulative distribution of the magnitude of a complex gaussian signal of equal power which lead to a Rayleigh law for its amplitude. As comparison, we have also reported a log-normal density which seems to have better behaviour than the Rayleigh law.

These curves indicate clearly that real speech spectral magnitude may have cumulative distribution that differs from the gaussian theoretical assumption of the MMSE model. These conclusions are also in agreement with those reported in the work of Porter and Boll [9] on optimal estimators for the restoration of noisy speech.

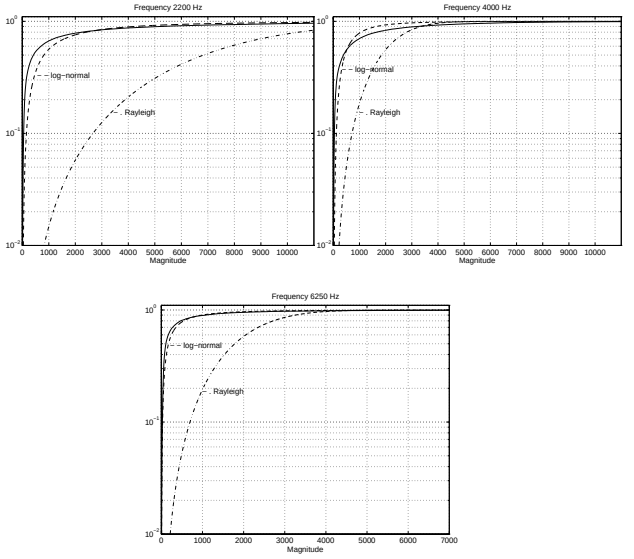

Figure 4. Spectral magnitude distribution of real speech (-), log-normal (- - -) and Rayleigh (- . -) distributions.

However despite these results on optimal behaviour of the MMSE and Wiener suppression rules, MMSE estimator has recently received much attention by many researchers for speech enhancement for hands-free mobile radio communications but also for the restoration of old recordings. How can we explain this fact ? An answer to this question can be found through the analysis of practical implementation of the Wiener and MMSE suppression rules.

4 PRACTICAL IMPLEMENTATION OF THE MMSE FILTER

The previous MMSE estimate has been derived under the implicit assumption that the a priori SNR and the noise power were known. However, in real implementation these parameters are unknown in advance and we have to provide estimators for these quantities. The following estimators are frequently used in real-time systems for respectively the noise power spectrum density, the a posteriori and a priori SNR :

$$\hat{P}_B^t(f_k) = \lambda \,.\, \hat{P}_B^{t-1}(f_k) + (1-\lambda).\left|B^t(f_k)\right|^2 \quad (5)$$

$$\hat{SNR}_{post}^t(f_k) = \frac{|X_k|^2}{\hat{P}_B^t(f_k)} \quad (6)$$

$$\hat{SNR}_{prio}^t(f_k) = (1-\beta).P\left[\hat{SNR}_{post}^t(f_k) - 1\right] + \beta.\frac{\left|\hat{S}^{t-1}(f_k)\right|^2}{\hat{P}_B(f_k)} \quad (7)$$

where $P[.]$ denotes half-wave rectification and the subscript $(.)^t$ holds for the actual time interval.

This last estimator has been first proposed in the original paper of Ephraim and Malah [4] and is evaluated in "decision-directed" approach since it takes into account the information in the current short-time frame but also the result of the processing in the previous frame. Furthermore, it has been reported [4,5] that this a priori SNR estimator acts as a key parameter in the reduction of speech distortions and musical noises. By incorporating these estimators in the previous definitions of the filters we are able to have practical implementation of the Wiener and MMSE suppression rules.

5 EXPERIMENTAL RESULTS

Experiments have been made with speech corrupted by background noise. The disturbing noise was recorded in car on a highway at 120 km/h speed and added to a clean speech signal recorded in a stopped car to obtain a noisy signal (see Figure 2(a) for spectrogram and time waveform). For practical implementation, we used 256 points (Fe = 8kHz) Fast Fourier Transforms of 16 ms hanning windowed signals. An additional constraint is added to the estimated time-variant impulse response of the noise reduction filter in order to respect the linear convolution operation. The noise power spectral density is evaluated during non speech activity periods with a first-order recursive filter (time constant 140 ms) according to equation 5.

Time evolution of the a priori SNR are given in figure 5 for the optimal value (equation 3) and its estimate (equation 7) for a frequency component of 1250 Hz. On low values of the signal to noise ratio, we can see that the time and frequency domain evolution of the internal parameters of the practical MMSE filter are sufficiently far from the optimal ones. We can notice that differences of approximately 20 dB are observed which induces a suppression factor far away from the optimal one thus introducing higher residual noise power. During speech activity periods, the estimated values are closed to the optimal ones but we can see that the a priori SNR estimation gives always lower values in comparison with the optimal ones. Listening tests confirm that audible distortions are still present in the processed speech

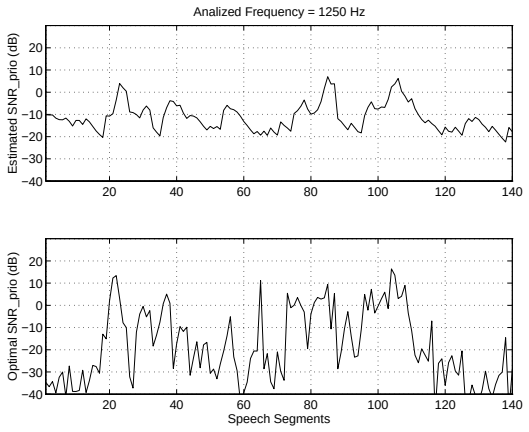

Figure 5. Optimal and estimated time evolution of the a priori SNR on real speech signal for a frequency of 1250 Hz.

The time evolution of the noise reduction factor is given in Figure 6 for the same period. During speech activity periods, we can see that using the optimal a priori SNR in the MMSE suppression rule results approximately in an additional noise reduction gain of 10 dB on the SNR at the output of the noise reduction filter. When the speech components are not present in the noisy signal, this additional noise reduction gain presents lower values of 5 dB.

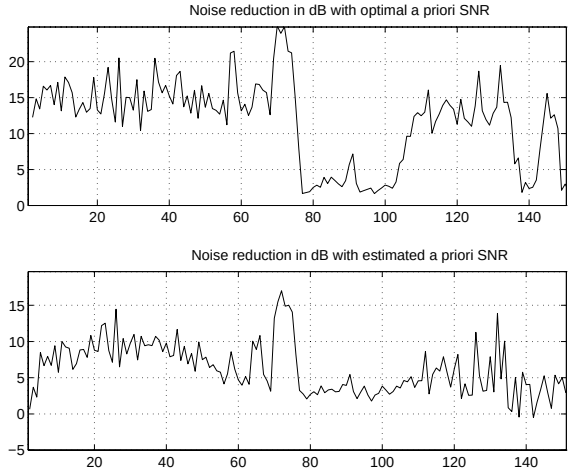

Figure 6 : Noise reduction in dB for practical and optimal MMSE filters.

Comparisons between optimal and practical behaviour of the MMSE filter clearly demonstrate that an effective improvement in the noise reduction process (i.e. quality of the enhanced speech, amount of noise reduction and nature of the residual noise) can be gained if greater attention is given to the estimation process of the filter parameters. Therefore, researchers have to focused on these estimates in order to provide a sufficiently high speech quality for these mobile services.

6 CONCLUSION

In this paper, the asymptotic performance of the MMSE Short-Time Spectral Estimator have been analysed and compared to the Wiener ones. We have shown that, when dealing with optimal behaviour of these filters, the Wiener suppression rule gives a high quality enhanced speech signal with significant noise reduction in comparison with the MMSE one. This result was explained by the fact that the Gaussian speech assumption in the MMSE model was not really true. Furthermore, we show that the practical implementation of the MMSE filter introduces a suppression factor far away from the optimal one. Therefore, we believe that better results can be gained by improving the filter parameters estimation in order to provide a sufficiently high speech quality for the mobile services.

ACKNOWLEDGEMENTS

The authors would like to thank CNPq and also Professor Jozue Geraldo CHIQUITO from Unicamp (Brazil).

REFERENCES

[1] Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", IEEE Trans. on ASSP, vol. 29, April 1979.

[2] Lim, Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", in Proc. IEEE, vol.67, Nº 12, December 1979.

[3] McAulay, Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. on ASSP, vol.28, Nº 2 April 1980.

[4] Ephraim, Malah, "Speech Enhancement Using MMSE Short-Time Spectral Amplitude Estimator", IEEE Trans. on ASSP, vol.32, Nº 6, Dec.1984.

[5] Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", IEEE Trans. on ASSP, April 1994.

[6] Yang "Frequency domain noise suppression approaches in mobile systems" in ICASSP, 1993.

[7] Brancaccio, Pelaez "Experiments on noise reduction techniques with robust voice detector in car environment" in Eurospeech , pp. 1259-1262, 1993.

[8] J. Hakkinen, M. Vaananen "Background noise suppressor for a car hands-free microphone" in 4th ICSPAT, pp. 300-307, Santa-Clara, California, 1993.

[9] Porter, Boll "Optimal estimators for spectral restoration of noisy speech" in ICASSP, 1984.

[10] Ephraim, Malah "Speech enhancement using optimal non-linear spectral amplitude estimator" in ICASSP, pp. 1118-1121, 1983.