SPEECH ENHANCEMENT BASED ON A PRIORI SIGNAL TO NOISE ESTIMATION

Pascal SCALART[1], *Jozue VIEIRA FILHO*[1,2,3]

[1]France Telecom - CNET/LAA/TSS/CMC, 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 speech enhancement in noisy environments. The main characteristics of available frequency domain noise reduction algorithms are firstly presented. We have confirmed that the A Priori SNR estimation leads to the best subjective results. According to these conclusions, a new approach is then developed which achieves a trade-off between effective noise reduction and low computational load for real-time operations. The obtained solutions demonstrate that subjective and objective results are much better than existing methods.

I - INTRODUCTION

The problem of enhancing speech degraded by uncorrelated additive noise, when only the noisy speech is available, has been widely studied in the past and it is still an active field of research. In many applications, such as mobile communication or speech recognition, efficient noise reduction techniques are needed, which require a low computational load. Many approaches have been investigated in that way. They include power spectral subtraction [1], Wiener filtering [2], soft-decision estimation [3] and Minimum Mean Square Error (or MMSE) estimation [4]. A common feature of the techniques presented in [1,2,3] is that the noise reduction process brings about very unnatural artifacts called "musical noise". For the MMSE approach, these residual noise characteristics have not been reported and our test confirmed these results.

In this paper, we first present a unified view of the main single microphone noise reduction techniques. In order to allow real time implementation of these algorithms and to achieve a trade-off between high quality noise reduction and low computational load, we propose to include the concept of A Priori SNR estimation in classical speech enhancement [1,3] schemes. Our results, based on classical objective measures and informal subjective tests, confirm the interest of this approach since the processed speech signals combine effective noise reduction with a highly reduced "musical noise" effect.

II - SPEECH ENHANCEMENT SYSTEMS

To date, conventional single microphone frequency domain speech enhancement techniques have been proposed on more or less an ad hoc basis. A common feature of these techniques is that the noise reduction process can be related to the estimation of a *Short-Term Suppression Factor.* Since the spectral components are assumed to be statistically independent, this factor is adjusted individually as a function of the relative local *A Posteriori Signal to Noise Ratio* on each frequency. In addition, a detector has to be used in order to determine whether the given noisy signal consists of noise only or speech plus noise, thus a binary model [3,4] which takes into account the uncertainty of speech presence in the noisy observations seems to be appropriate.

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-stationnarity of the speech signal is guaranteed over the period T. It is useful to consider the amplitude estimate $\hat{A}_k$ as being obtained from X_k by a multiplicative non-linear gain function defined by $G(f_k) \underline{\underline{\Delta}}\, \hat{A}_k \,/\, X_k$. In order to present a unified view of single microphone noise reduction techniques, we can express, without loss of generality, the optimal gain function as the product of the standard gain G_0 by a term which contributes to the "soft-decision" aspect of the estimate as given by :

$$G(f_k) \;=\; \frac{\Lambda(X_k,q_k)}{1+\Lambda(X_k,q_k)}\;\; G_0(f_k) \qquad (1)$$

where $\Lambda(X_k, p_k)$ is the generalized likelihood ratio taking into account the uncertainty of speech presence in the noisy observations defined by :

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

with $\mu_k \,\underline{\underline{\Delta}}\, (1-q_k)/q_k$, where q_k is the probability of signal absence in the *k*th spectral component, and *p(.)*

0-7803-3192-3/96 $5.00©1996 IEEE

denotes a probability density function. H_k^0 and H_k^1 denote the two hypotheses of signal absence and presence, respectively, in the kth spectral component. Note that if $q_k = 0$, $\Lambda / (1+\Lambda)$ equals unity, and $G(f_k)$ turns out to be equal to the standard gain function $G_\Diamond$ when the speech signal is always present in the noisy observation.

Defining local *A Posteriori* and *A Priori* SNRs by :

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

existing methods including power estimation, maximum likelihood estimate, Wiener estimation soft-decision method and MMSE estimate can be related to equation (1) as shown in Table 1.

These amplitude estimates (3) have been derived under the implicit assumption that the A Priori SNR and noise spectral density function are known. However, in practical implementations of speech enhancement systems, these parameters are unknown in advance as the noisy speech alone is available. Moreover, it has been reported [4,5] that the A Priori SNR acts as a key parameter (rather than the noise variance) in the reduction of speech distortions and musical noises. In order to have real-time algorithms, two main directions can then be found in the literature : the first one tries to defined from informal listening a fixed (generally 5 or 7 dB) optimal value for the *A Priori* SNR [6,7]. The second one [4] replaces the corresponding unknown parameters by the following estimates for respectively the noise power spectral density, the A Posteriori and A Priori SNRs :

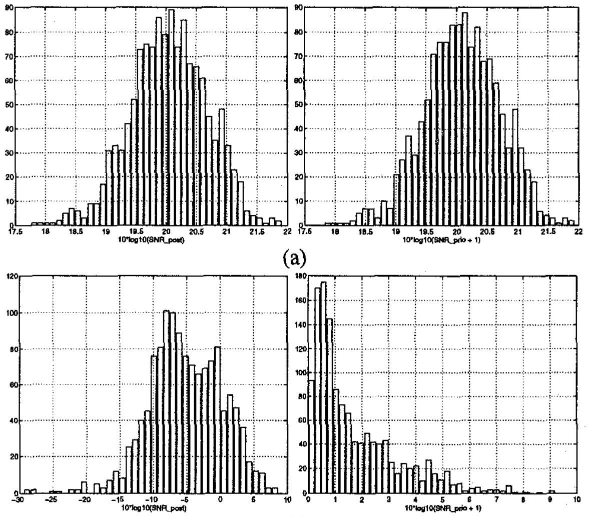
$$\hat{P}_B^t(f_k) = \lambda . \hat{P}_B^{t-1}(f_k) + (1-\lambda).\left|B^t(f_k)\right|^2 \tag{4}$$

$$S\hat{N}R_{post}^t(f_k) = \frac{|X_k|^2}{\hat{P}_B^t(f_k)} \tag{5}$$

$$S\hat{N}R_{prio}^t(f_k) = (1-\beta).P\left[S\hat{N}R_{post}^t(f_k) - 1\right] + \beta.\frac{\left|\hat{S}^{t-1}(f_k)\right|^2}{\hat{P}_B(f_k)} \tag{6}$$

where *P[.]* denotes half-wave rectification and the subscript $(.)^t$ is for the actual time interval . We can notice that the MMSE algorithm [4] has recently received much attention by many researchers. However, it is very hard to show if the good behaviour of Ephraim and Malah's algorithm comes from this "decision-directed" estimate or if it comes from the gaussian statistical model coupled with the MMSE approach? If we look at Porter and Boll results [9] (confirmed by our own investigations) on cumulative distribution of real speech spectral magnitudes, it could be possible to consider the first hypothesis.

Theoretical comparisons between the recursive estimate (6) and its true value given in (3) are very complicated due to its highly non-linear nature. However, in order to show the superiority of this kind of estimations, we present on Figure 1 the histograms obtained on 1kHz sinusoidal test signal corrupted by gaussian additive white noise. We can notice on Figure 1.a that when the local SNR on the 1kHz component is high, the two SNR estimates give approximately the same mean and standard deviation. When the SNR comes low (Figure 1.b), the A Posteriori SNR exhibits very low values below 0 dB and thus large standard deviation. On the other hand, the A Priori SNR estimator (evaluated from (6) with β = 0.98) still have a good behaviour with low standard deviation. However, we can notice a bias on this last estimator (the true SNR is 4 dB) when the SNR is low. This observation can be partly explained by the arbitrarily zeros values given by half-wave rectification in (6) when $\mathrm{SNR}_{\mathrm{post}}$ is negative. We believe, as in [10], that better results can be obtained by improving the Priori SNR estimation (a possible approach has been proposed in [8]).



(a)

(b)

Figure 1 : Histograms of the A Priori and A Posteriori SNR estimates in dB (a) true SNR = 20 dB and (b) true SNR= 4 dB.

According to these results, we propose to include the concept of A Priori SNR in classical speech enhancement schemes such as Wiener, spectral subtraction, or Maximum Likelihood estimates (see Table 1). This can be done by considering $E\{SNR_{post}(f_k)\} = 1 + SNR_{prio}(f_k)$ leading to :

$$G_\Diamond^{PE} = \sqrt{\frac{SNR_{prio}(f_k)}{1+SNR_{prio}(f_k)}} \tag{7}$$

$$G_\Diamond^{W} = \frac{SNR_{prio}(f_k)}{1+SNR_{prio}(f_k)} \tag{8}$$

$$G_\Diamond^{ML} = \frac{1}{2}\left[1+\sqrt{\frac{SNR_{prio}(f_k)}{1+SNR_{prio}(f_k)}}\right] \tag{9}$$

where the A Priori SNR estimate is evaluated from the "decision-directed" approach (6). We can also notice that the proposed estimators give low computational load for real time operations.

III - EXPERIMENTAL RESULTS

Experiments have been made with speech corrupted by background noise. The disturbing noise was recorded in vehicle 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 for spectrogram and time waveform). For practical implementation, we used 512 points (Fe = 16kHz) Fast Fourier Transforms of 32 ms hanning windowed signals. The noise power spectral density is evaluated during non speech activity periods with a first-order recursive filter (time constant 140 ms).

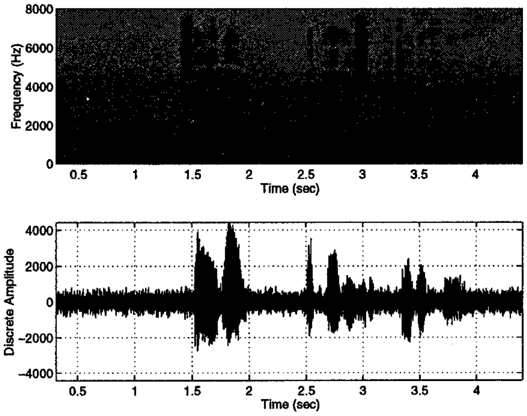


Figure 2 : Spectrogram of the noisy speech signal.

Due to space limitations, only results obtained from the Power Spectral Subtraction technique (where musical noise is more annoying) will be presented. The parameters of the classical and proposed algorithms are chosen in order to give the same average noise power reduction (≈10 dB) during non-speech activity periods. The corresponding signals processed by classical method and the proposed algorithms are presented respectively on Figures 3 and 4.

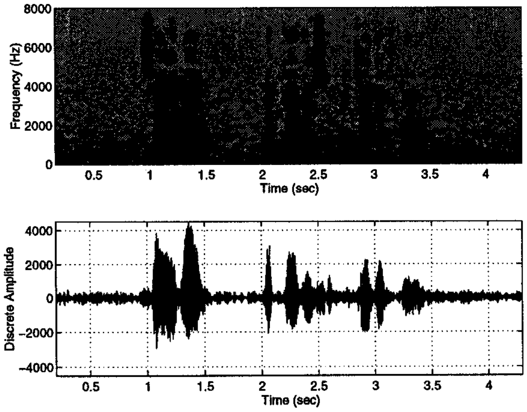


Figure 3 : Spectrogram of the estimated speech signal with conventional Spectral Subtraction method.

These figures clearly demonstrate that an effective noise reduction can be gained during non speech activity periods. However, by careful examination of the spectrograms during these periods, we can notice spectral artifacts on Figure 3 resulting from random tone bursts. This is not the case on Figure 4.

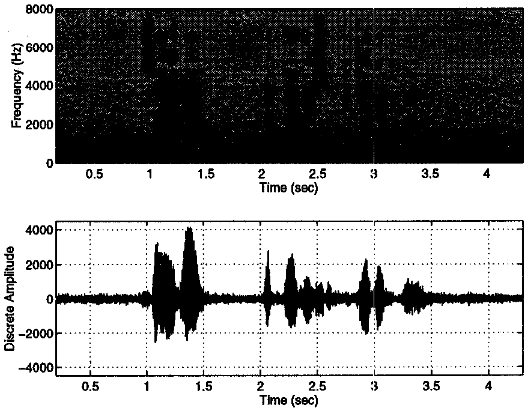


Figure 4 : Spectrogram of the estimated speech signal with the proposed Spectral Subtraction method.

This "musical noise" phenomena can also be analyzed on Figure 5 where the histograms of the 1500Hz residual noise power component are presented for classical and proposed methods. We can see that the power distribution of the 1500Hz residual noise component is more regular with the proposed method and that this last algorithm gives also a lower standard deviation. These two characteristics introduce, for the proposed method, a more uniform residual noise power spectral density without "musical-noise" effect.

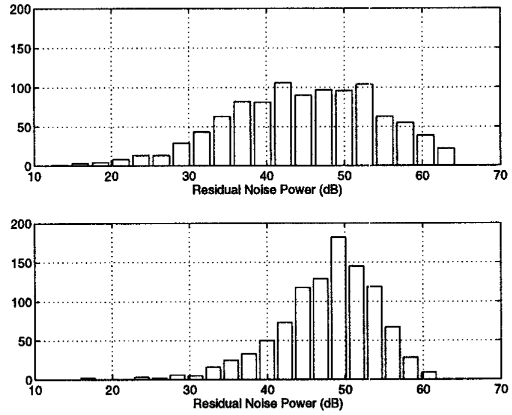


Figure 5 : Histograms of the residual noise power (in dB) for classical (upper) and proposed (lower) method.

In order to provide information on speech distortions, we have chosen to represent the distortions which result in vocal tract spectral mismatch though the objective cepstral distance which provides a quantifiable mean of assessing distortions. The cepstral distances between the enhanced speech signal components and the noise-free speech signal are represented on Figure 6. We can see that the proposed

Methods	Noise suppression function	Probability and likelihood ratio
Power Estimation [1,2]	$G_{\Diamond}^{PE} = \sqrt{\frac{SNR_{post}(f_k)-1}{SNR_{post}(f_k)}}$	$q_k = 0$
Maximum Likelihood Estimate [3]	$G_{\Diamond}^{ML} = \frac{1}{2}\left[1+\sqrt{\frac{SNR_{post}(f_k)-1}{SNR_{post}(f_k)}}\right]$	$q_k = 0 \Rightarrow \quad \Lambda(X_k, p_k) = 1$
Wiener Estimate [2]	$G_{\Diamond}^{W} = \frac{SNR_{post}(f_k)-1}{SNR_{post}(f_k)}$	$q_k = 0 \Rightarrow \quad \Lambda(X_k, p_k) = 1$
McAulay&Malpass Estimate [3,6,7]	$G_{\Diamond}^{MA} = \frac{1}{2}\left[1+\sqrt{\frac{SNR_{post}(f_k)-1}{SNR_{post}(f_k)}}\right]$	q_k=0.5, $\Lambda = \exp(-SNR_{prio})\, I_0\left[2\sqrt{SNR_{prio}\, SNR_{post}}\right]$
MMSE Estimate [4,5,8,10]	$G_{\Diamond}^{EM} = \frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{SNR_{post}}\left(\frac{SNR_{prio}}{1+SNR_{prio}}\right)} \times F\left[SNR_{post}\left(\frac{SNR_{prio}}{1+SNR_{prio}}\right)\right]$	$0 < q_k < 1$, $\Lambda = \mu_k \frac{\exp\left(SNR_{post}\, SNR_{prio} / 1+SNR_{prio}\right)}{1+SNR_{prio}(f_k)}$

Table 1 : Conventional single microphone speech enhancement methods.

algorithm introduces less distortions during speech activity periods but also during non-speech activity periods (the noise characteristics are kept unchanged excepted for the averaged noise power).

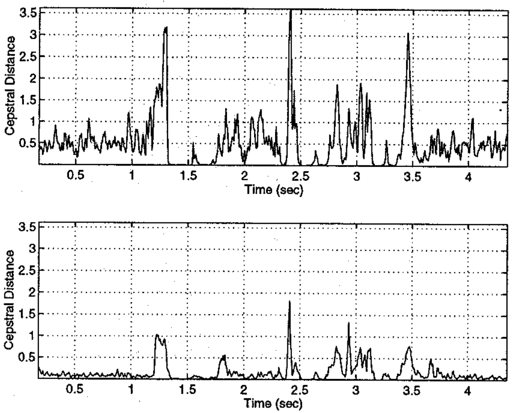


Figure 6 : Cepstral distance for classical (upper) and proposed (lower) methods.

Informal subjective tests confirm these results and show that the proposed algorithm produces perceptually more pleasant enhanced speech signals.

IV - CONCLUSION

In this paper, a unified view of the main single microphone noise reduction techniques in the frequency domain has been presented. We have proposed a new approach for the speech enhancement methods based on the use of the A Priori Signal to Noise Ratio. These new solutions allow significant noise power reduction without introducing "musical-noise" effects. Results show that obtained solutions demonstrate much better subjective results than existing methods. Moreover, these methods combine effective noise reduction and low computational load for real-time operations.

ACKNOWLEDGEMENTS

The authors would like to thank CNPq and also Professor Jose 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, N° 12, December 1979.

[3] McAulay, Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. on ASSP, vol.28, N° 2 April 1980.

[4] Ephraim, Malah, "Speech Enhancement Using MMSE Short-Time Spectral Amplitude Estimator", IEEE Trans. on ASSP, vol.32, N° 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.