ANALYSIS OF BACKGROUND NOISE REDUCTION TECHNIQUES FOR ROBUST SPEECH CODING

David Virette[1], Pascal Scalart[1,2], Claude Lamblin[1]

[1] FRANCE TELECOM R&D, 2. Av. Pierre Marzin, 22307 Lannion Cedex, France
[2] LASTI – Ecole Nationale Supérieure des Sciences Appliquées et de Technologie, 22305 Lannion Cedex, France
{david.virette, pascal.scalart, claude.lamblin@rd.francetelecom.com}

ABSTRACT

In general, low rate speech coding systems do not have their own mechanism to reduce background noise from the speech signal. This is due to the complexity of the speech signal and limitations in the scope of many speech coding systems. As a consequence, most speech enhancement systems to date have attempted to process the speech waveform directly and independently from the speech coding system, before the encoding of the speech signals. In this paper, we propose several methods to optimize speech enhancement techniques in order to improve the estimation of the CELP speech codec parameters (short-term and long-term parameters). Experimental results for two speech enhancement techniques are presented in conjunction with the ETSI AMR speech codec.

1. INTRODUCTION

Single-channel noise reduction is a quite difficult challenge, since the speech and the noise are mixed within the same observation signal, *i.e.* $y(t)=s(t)+n(t)$.

Improving the speech quality remains today a major challenge in a competitive field such as telecommunications. To achieve good performance of coding in noisy environments, several studies have been recently proposed for the definition of an optimized solution to a wider problem: jointly optimizing the noise reduction processing and the source encoding operations, and thus to a point where these two problems are no longer considered as independent. Several prospective studies have proposed the combination of a noise reduction system as a pre-processing unit in tandem with a low-rate speech coder [1, 2, 3]. Other proposals were also made in the field of standardization like the ETSI Adaptive Multi-Rate (AMR) [4] or the ITU-T 4 kbit/s coders [5]. Other works [6] are related to an optimized procedure to estimate the LP parameters (in the LSF domain) for noisy speech coding.

In this paper we provide an experimental analysis of the influence of the combination of a noise-reduction pre-processing unit in tandem with a low-rate CELP speech coder on the speech quality. We first describe in Section 2 the speech enhancement techniques that have been considered throughout this study. After briefly reviewing the main characteristics of CELP speech coders, we compare in Section 3 the performance of different noise pre-processing units on the estimation of the short-term and long-term parameters of the NB-AMR speech coder. All the investigations presented in this paper are based upon the NB-AMR coder. However, the conclusions of this study can be applied to a wide range of speech CELP coders such as the ITU-T G.729 [7] or the ITU-T G.723.1 [8].

2. NOISE REDUCTION PRINCIPLES

In this section, we first recall the main characteristics of two specific suppression rules, which will be used throughout this paper. These techniques are based on the short-time spectral attenuation principle, which can be considered as the predominant approach for speech enhancement due to its simplicity of implementation and applicability to various noise environments.

2.1. Basic Suppression Rules

The Wiener suppression rule provides the optimal linear estimator (in the minimum mean-square error (MMSE) sense) of the *k*th signal spectral component given the noisy observation signal $y(t)=s(t)+n(t)$. Moreover, assuming statistical independence between noise and speech processes and between frequency bins, the short-time Wiener amplitude estimator is obtained by a multiplicative nonlinear gain function which is defined by

$$G_{\mathrm{MMSE}}(p,w_k)=\mathcal{SNR}_{\mathrm{prio}}(p,w_k)\Big/\left[1+\mathcal{SNR}_{\mathrm{prio}}(p,w_k)\right] \quad (1)$$

where p and w_k stand for the time and frequency indices, respectively. This spectral gain depends on a single parameter, which is interpreted as the *a priori signal-to-noise ratio* at frame p defined by

$$\mathcal{SNR}_{\mathrm{prio}}(p,w_k)=\left|S(p,w_k)\right|^2\Big/\gamma_n(w_k) \quad (2)$$

where $\gamma_n(w_k)$ denotes the noise power at frequency w_k.

In the second approach considered in this paper, the noise reduction filter is derived under an assumed Gaussian statistical model as proposed by Ephraïm and Malah in [9]. In this case, the MMSE short-time spectral amplitude (STSA) estimator is obtained as a spectral gain $G_{STSA}(p,w_k)$ that is applied to each spectrum value $Y(p,w_k)$ of the noisy speech frame; it is given by

$$G_{\mathrm{STSA}}(p,w_k)=\frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{\mathcal{SNR}_{\mathrm{post}}(p,w_k)}\right)\left(\frac{\mathcal{SNR}_{\mathrm{prio}}(p,w_k)}{1+\mathcal{SNR}_{\mathrm{prio}}(p,w_k)}\right)}\times\mathcal{M}\left[\mathcal{SNR}_{\mathrm{post}}(p,w_k)\left(\frac{\mathcal{SNR}_{\mathrm{prio}}(p,w_k)}{1+\mathcal{SNR}_{\mathrm{prio}}(p,w_k)}\right)\right] \quad (3)$$

where $\mathcal{M}$ stands for the function

$$M[x]=\exp(-x/2)\left[(1+x)I_0(x/2)+x\,I_1(x/2)\right] \quad (4)$$

where I_0 and I_1 are the modified Bessel functions of the first kind of zero and first order, respectively. In (3), the *a posteriori signal-to-noise ratio* is defined by

$$\mathcal{SNR}_{\text{post}}(p,w_k) = \max\left\{0, \left|Y(p,w_k)\right|^2 / \gamma_n(w_k) - 1\right\} \quad (5)$$

and expresses the instantaneous $\mathcal{SNR}$ in frame p for each spectral component (estimated from the power-subtraction method).

2.2. Taking into account speech uncertainty

The previous suppression rules have been derived under the assumption of signal presence in the noisy observation. However, signal absence in the noisy observation is frequent since speech signals contain large portions of silence. The above discussion suggests the following statistical model where a statistically independent random appearance of the signal in the noisy spectral components is assumed. Based on such a model, the « soft-decision » MMSE-STSA estimator that takes into account the uncertainty of signal presence in the noisy observation is given by [9].

$$G^{\text{SD}}_{\text{STSA}}(p,w_k) = \frac{\Lambda(p,w_k)}{1+\Lambda(p,w_k)} G_{\text{STSA}}(p,w_k)\Big|_{\mathcal{SNR}_{\text{prio}} = \frac{|S(p,w_k)|^2}{(1-q_k)\gamma_n(w_k)}} \quad (6)$$

where the generalized likelihood ratio is defined by

$$\Lambda(p,w_k) = \frac{1-q(w_k)}{q(w_k)} \frac{\exp\left(\mathcal{SNR}_{\text{post}}\left[\frac{\mathcal{SNR}_{\text{prio}}}{1+\mathcal{SNR}_{\text{prio}}}\right]\right)}{1+\mathcal{SNR}_{\text{prio}}} \quad (7)$$

where p and w_k have been omitted for compactness reasons, and where $q(w_k)$ is the signal absence probability in the *k*th spectral component. In practice, this parameter is classically set to a value of 0.2 for each frequency bin, but this probability can also be made time-varying $\left[i.e.\ q(p,w_k)\right]$ as proposed in [10] and [11].

Following the same analysis as it was done previously, the optimal Wiener filter under speech presence uncertainty was introduced in [12] as

$$G^{\text{SD}}_{\text{MMSE}}(p,w_k) = \frac{\Lambda(p,w_k)}{1+\Lambda(p,w_k)} \frac{\mathcal{SNR}_{\text{prio}}(p,w_k)}{1+\mathcal{SNR}_{\text{prio}}(p,w_k)} \quad (8)$$

where $\Lambda(p,w_k)$ is defined by (7).

In the previous relations, the *a priori signal-to-noise ratio* is the dominant parameter (see [13] for a discussion on this subject). However, this parameter is unknown since it represents the information on the unknown spectrum magnitude. In practice, this $\mathcal{SNR}$ is estimated in a « decision-directed » approach as the average $\mathcal{SNR}$ to exploit the local stationarity of the speech signal

$$\hat{\mathcal{SNR}}_{\text{prio}}(p,w_k) = \alpha \left|\hat{S}(p-1,w_k)\right|^2 / \gamma_n(w_k) + (1-\alpha)\mathcal{SNR}_{\text{post}}(p,w_k) \quad (9)$$

The choice of the mixing parameter α is guided by a trade-off between the degree of smoothing and the acceptable level of transient distortion brought to the signal. Simulations show that, in order to reduce the musical noise effect, one will choose value of α as close to one as possible.

Two noise reduction systems will be considered in this study: the soft-decision Wiener estimator [using (5), (7) , (8) and (9)] and the soft-decision MMSE-STSA estimator [using (5), (6), (7) and (9)]. For both systems, the signal absence probability in the *k*th spectral component is set to 0.2 and the mixing parameter α is set to 0.98.

3. INFLUENCE OF NOISE PRE-PROCESSING ON SPEECH CODERS

In this section, we examine the influence of the background noise suppressors on the performance of CELP speech codecs. We present the different coders that have been considered throughout this study and propose several methods to optimize speech enhancement techniques in conjunction with these codecs.

3.1. Selected speech coders

It is known that the coding of noisy speech becomes significantly more difficult as bit rates are decreasing. In this study we consider low bit-rate CELP speech coders which have been proposed in new standards for mobile or voice over packet switched networks: the ITU-T G.729 [7], the ITU-T G.723.1 [8], and the ETSI narrow-band AMR [4] speech codec. As several modes are available for each of these codecs, we have selected the following bit rates: the G.723.1 at 5.3 kbit/s, the G.729 at 6.4 kbit/s, and the NB-AMR at 5.15 kbit/s. The study has been performed over several CELP coders. However, for sake of compactness, only the results for the NB-AMR speech codec will be reported since we found similar results for the other coders.

3.2. Implementation and Performance Evaluation

The input noisy speech is first windowed by a half-overlapped Hamming window of length 256 points and then spectrally decomposed by the fast Fourier transform (FFT). The spectral amplitude of the noise-suppressed speech at frequency w_k is estimated either by the soft-decision Wiener estimator or by the soft-decision MMSE STSA estimator, as previously mentioned. The spectral amplitude is then combined with the noisy phase. Synthesis of the pre-processed signal is realized by applying an inverse FFT on the resulting spectrum, and by overlapping and adding two consecutive frames according to the weighted overlap-add method.

In our experiments, we have also considered two additional estimators corresponding to the previous ones (Wiener and MMSE-STSA with speech uncertainty) under the knowledge of the optimal values of the *a priori* $\mathcal{SNR}$ given by (2) [in place of (9)]. The motivation for these two additional noise reduction schemes is to provide an absolute quality reference for the enhanced speech signals.

For objective evaluation, we generated noisy speech data by adding noises (either car, babble or street noises) to clean speech files. The global SNR of the noisy speech signals is varying between -10 and 40 dB. To obtain a fair comparison between the proposed algorithms, the voice and noise-only regions are obtained using manually marked boundaries and are the same for every algorithm.

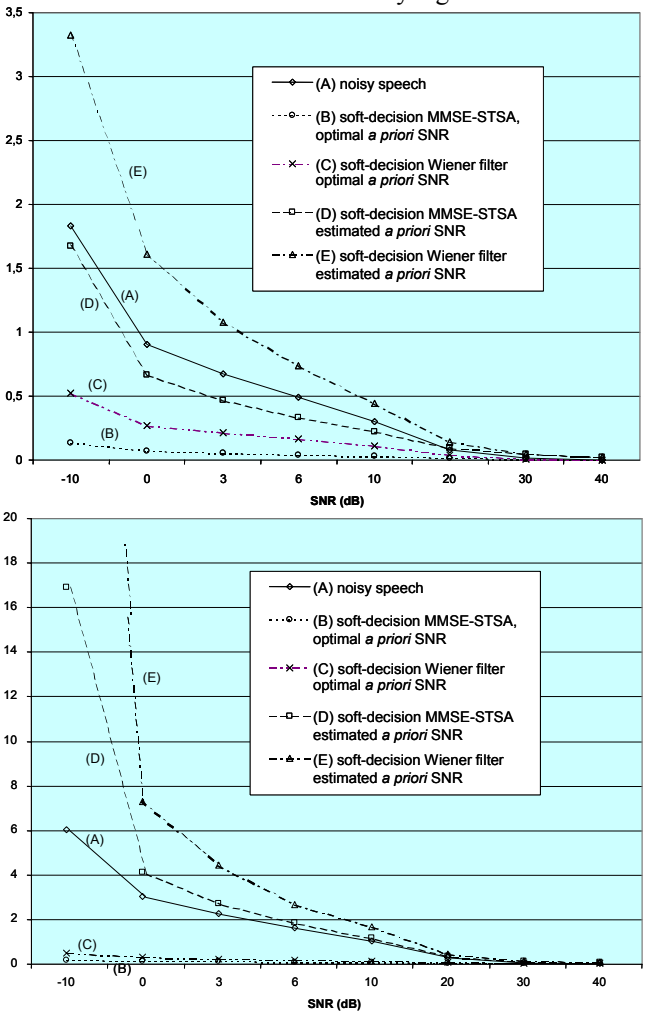

Fig. 1. Itakura-Saïto distance measure values between the original clean speech and the enhanced speech files [(A) noisy speech, (B) MMSE-STSA optimal, (C) Wiener optimal, (D) MMSE-STSA estimated, (E) Wiener estimated] for car noise (upper) and babble noise (lower).

3.3. Short-Term (LPC) parameters

To provide a measure of spectral distorsion introduced on the LPC spectrum envelope by the noise suppression methods, we compare in Figure 1 the Itakura-Saïto distance evaluated between the original clean speech and the enhanced speech for the four different noise reduction schemes (curves B, C, D, and E). To compute this distance, the LPC coefficients (of the clean speech and the enhanced speech) are extracted from the 10^{th} order linear prediction filter (done once per 20 ms frame) of the NB-AMR codec. Also shown in this figure is the distance measure evaluated between the noisy (*i.e.* with no pre-processing - curves A) and the clean speech files. At very high SNRs, it can be easily shown that the four methods have approximately the same behavior. As the SNR decreases, major differences can be seen when the global SNR is lower than 20 dB.

Let us first consider in Figure 1 the asymptotic bounds of noise reduction associated with the definition of ideal system performance (curves B and C). For both optimal systems, it is clearly shown that the short-term prediction parameters of the enhanced speech signals (coded and decoded - curves A) are close to the ones of the original clean speech (coded and decoded). These experiments tell us the following conclusion: *Ideally, the short-term parameters of the original clean speech can be recovered at the output of any CELP speech coder by inserting a pre-processing noise reduction unit based on the short-term spectral attenuation principle (i.e. by estimating just the spectral magnitude and leaving the phase as it is)*. Therefore, the main issue is to estimate properly the *a priori* $\mathcal{SNR}$ given by (2) which can be considered as the main parameter of the noise reduction system.

In practice ideal noise reduction cannot be achieved since the *a priori* $\mathcal{SNR}$ is not known but it has to be estimated (see Figure 1, curves D an E). As a consequence, due to short-time instationnary of random noise components the quality of the processed signal rapidly decreases when the noise power increases. Moreover, when considering non-stationary noise sources such as babble noise, the Itakura-Saïto performance using a noise reduction as pre-processing give worse results than in the no-noise reduction case. To improve robustness to the noise types, we found two possible solutions: a limitation in the length of the impulse response of the noise reduction filter, or a limitation in the maximum value of short-term spectral attenuation provided by the noise reduction filter.

3.4. Long-Term (LTP) parameters

To evaluate the influence of the noise reduction schemes on the long-term parameters estimation of the coded speech signals, we compare in Figure 2 the candidate delays of the clean and enhanced speech signals obtained from the AMR-NB open-loop pitch analysis. In this configuration, the noise reduction system is inserted as a pre-processing unit for the open-loop pitch search. To take into account the pitch lag sensitivity, a 5% error (*i.e.* ±1 for an open-loop pitch value of 20, and ±7 for a value of 140) is allowed around the value of the candidate delay obtained from an open-loop pitch analysis in the noise-free configuration with no preprocessing unit.

From Figure 2, in the car noise case (stationary) the better performance obtained when using a noise reduction pre-processing unit (either optimal or estimated) in comparison with the no-noise reduction case can be easily observed. For non-stationary noises such as babble noise, we can see in Figure 2 that the different curves are close. In comparison with the noisy speech case, enhanced performance is obtained only if ideal noise processing is used (either MMSE or Wiener).

To further analyze the effect of the noise-suppressor pre-processing unit, we analyze in Figure 3 the relative error in

the delays in the open-loop pitch analysis defined by $\left(P_{\text{clean}} - \hat{P}_{\text{enhanced}}\right)\Big/P_{\text{clean}}$. To compute this histogram, only speech frames for which the normalized autocorrelation in the open-loop pitch analysis procedure is greater than a fixed threshold have been considered (voiced frames). We can see that the noise suppression unit avoids selecting pitch multiples (up to 3 or 4 times the pitch delay). These results demonstrate the practical interest of inserting specific mechanism (such as noise reduction pre-processing in the open-loop pitch analysis) into the core processing of low-rate speech codecs.

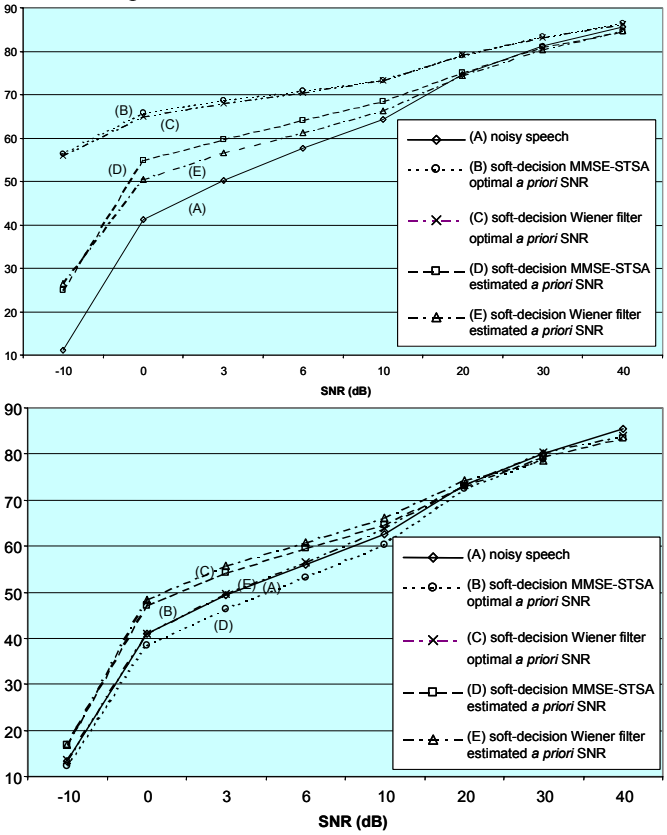

Fig. 2. Percentage of correct value of the candidate delay obtained from the NB-AMR open-loop pitch analysis in comparison with the noise-free configuration with no pre-processing unit: car (upper) and babble (lower) noises.

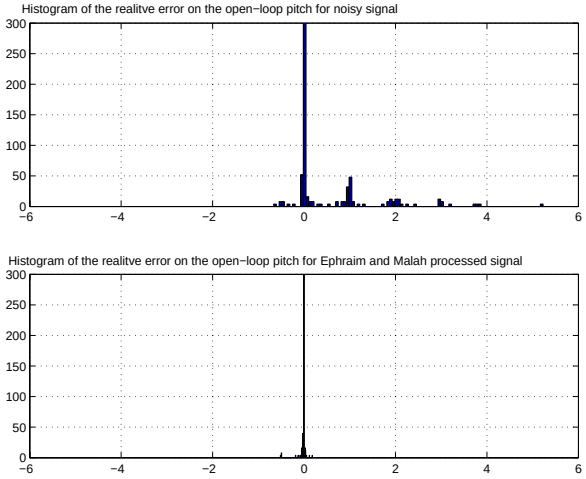

Fig. 3. Relative error in the open-loop pitch delay (car).

4. CONCLUSIONS

In this paper we show how the performance of a joint speech enhancement and coding system can be improved by inserting noise reduction pre-processing units dedicated to the estimation of the short-term and/or long-term parameters of low-rate CELP codecs in the presence of background noise. We show that such an optimized system gives improved results in conjunction with the ETSI NB-AMR codecs even for low SNR conditions. Results are applicable to a wide range of speech coders since we made extensive experiments with other speech coders such as the ITU-T G.729 and G.723.1. The results suggest the use of different speech enhancement techniques as pre-processors for different parameter extraction modules of the coder, since different modules make use of different aspects of the input speech in order to encode it. In that sense, we agree with the spirit of the IS-641 speech coder (at 7.4 kbit/s ACELP codec), and with the work of Accardi and Cox [1].

5. REFERENCE

[1] A. J. ACCARDI, R. V. COX, "A modular approach to speech enhancement with an application to speech coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, #2099, 1999.

[2] G. GUILMIN, *et al.*, "Study of the influence of noise pre-processing on the performance of a low bit rate parametric speech coder," in *Proc. European Conf. Speech and Signal Processing*, pp. 2367-2370, Budapest, Hungary, 1999.

[3] R. MARTIN, R. V. COX, A. ACCARDI., "Low delay analysis/synthesis schemes for joint speech enhancement and low bi rate speech coding," in *Proc. European Conf. Speech and Signal Processing*, 1999.

[4] ETSI GSM 06.90., "Digital cellular telecommunications system: Adaptive Multi-Rate speech transcoding," 01/1999.

[5] J. THYSSEN *et al.*, "A candidate for the ITU-T 4kbit/s speech coding standard," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, #2336, 2001.

[6] R. MARTIN, I. WITTKE, P. JAX., "Optimized estimation of spectral parameters for the coding of noisy speech ," in *Proc.Int. Conf. Acoutics, Speech, and Signal Processing*, 2000.

[7] ITU-T REC. G.729., "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction," 1996.

[8] ITU-T DRAFT REC. G.723.1., "Dual rate speech coder for multimedia telecommunication transmitting at 5.3 & 6.3 kbit/s," 1996.

[9] Y. EPHRAIM, D. MALAH, "Speech enhancement using a minimum mean-square error short time amplitude estimator," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[10] I. YANN SOON *et al.*, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," in *Signal Processing 75*, pp. 151-159, 1999.

[11] D. MALAH *et al.,* "Study of the influence of noise pre-processing on the performance of a low bit rate parametric speech coder," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1999.

[12] A. AKBARI AZIRANI, *et al.*, "Speech enhancement using a Wiener filtering under signal presence uncertainty," in *Proc. European Conf. Speech and Signal Proc.*, pp. 971-974, 1996.

[13] O. CAPPE, "Elimination of the musical noise phenomenon," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 345-349, April 1994.