ON THE OPTIMISATION OF NOISE PRE-PROCESSING FOR A LOW BIT RATE PARAMETRIC SPEECH CODER

Gwénaël GUILMIN1,2, Régine LE BOUQUIN JEANNES1, and Philippe GOURNAY2

1Laboratoire de Traitement du Signal et de l’Image
Université de Rennes 1
Bât. 22- Campus de Beaulieu
35042 Rennes CEDEX – France
2Thomson-CSF Communications
66, rue du fossé blanc – B.P. 156
92231 Gennevilliers CEDEX – France
e_mail: gwenael.guilmin@univ-rennes1.fr

ABSTRACT

This paper presents a study of the influence of a single-sensor noise pre-processing method (i.e. applied before coding) on the performance of a parametric low bit rate speech coder. The speech coder we consider estimates four parameters: the fundamental frequency, a voicing information, a set of linear prediction coefficients, and the energy. We first study the influence of background noise and noise pre-processing on those parameters and on the global output intelligibility. Then, we propose two new noise pre-processing methods based on the combination of two algorithms. Finally, we present some tests results that underline the advantages gained from using a combined pre-processing technique.

Keywords: noise reduction, low bit rate speech coding, intelligibility.

1. INTRODUCTION

Since the first well-known US and NATO LPC10 standards, the performance of low bit rate speech coders (below 4.8 kbits/s) improved a lot. New vocoders, such as IMBE, PWI, MELP and HSX, provide increased intelligibility and quality, but are still very sensitive to the acoustic environment, and especially to additive background noise. At the same time, particular attention was focused on speech enhancement in order to improve the intelligibility and the quality of speech degraded by additive background noise. Spectral subtraction and Wiener filtering now represent the most popular methods for single-sensor noise reduction [1][2][3]. As a consequence of those parallel developments, the future low bit rate speech coding standard established by NATO will very likely include a noise reduction procedure (this work is currently in progress). Therefore, it seems very important to optimize the association between the noise pre-processing and the low bit rate speech coder. The aim of this paper is to propose a noise reduction approach that is optimized for a low bit rate parametric speech coder. This kind of association is very interesting for any kind of low bit rate communication, including internet protocol telephony. The noise reduction system under consideration (namely Wiener filtering under signal presence Uncertainty - WU) and the parameters extraction procedure of the speech coder are presented in section 2. The influence of various types and levels of noise on the performances of the coder is studied in section 3. We used both objective measures based on the estimated speech parameters, and subjective intelligibility evaluations. The advantages gained from using WU - but also the shortcomings of this method - are particularly pointed out. In section 4, we propose two new noise pre-processing methods based on a combination of two basic noise reduction systems that overcome most of the limitations of the WU method. The improvement in objective and subjective performance brought by these new methods is studied in section 5.

2. DESCRIPTION OF THE NOISE PRE-PROCESSING METHOD AND OF THE LOW BIT RATE PARAMETRIC SPEECH CODER

2.1 Wiener filtering under signal presence uncertainty

The noise reduction system we use is a modified Wiener filtering [4]. This algorithm is based on Ephraim and Malah estimator [5], given by the
minimization of the mean-square spectral error. Two hypotheses are considered: presence and absence of speech signal. Let \( x(t), s(t) \) and \( n(t) \) be respectively the degraded signal, the original speech signal and the additive background noise, and \( X_{k,m}, S_{k,m} \) and \( N_{k,m} \) their \( k \)th spectral components obtained by short-term discrete Fourier transform (STDFT), for each frame \( m \) of 180 points:

\[
X_{k,m} = S_{k,m} + N_{k,m}. \tag{1}
\]

The filter \( H_{k,m} \), performed on \( X_{k,m} \) to give the estimate \( \hat{S}_{k,m} \), is defined by:

\[
H_{k,m} = W_{k,m} G_{k,m} \tag{2}
\]

where \( W_{k,m} \) corresponds to Wiener filtering and \( G_{k,m} \) represents the uncertainty of speech signal presence, as defined by Ephraim and Malah [6] with an estimation of a priori and a posteriori signal to noise ratios (\( SNR_{\text{prior}}^{\text{SNR}} \) and \( SNR_{\text{post}}^{\text{SNR}} \) respectively). We can express the Wiener filter by:

\[
W_{k,m} = \frac{SNR_{k,m}^{\text{prior}}}{SNR_{k,m}^{\text{prior}} + 1} \tag{3}
\]

and the term of uncertainty by:

\[
G_{k,m} = \frac{\Lambda_{k,m}}{\Lambda_{k,m} + 1} \tag{4}
\]

where \( \Lambda_{k,m} \) corresponds to the generalized likelihood ratio defined in [5] using the Gaussian statistical model assumed for the spectral components.

In our application, \( SNR_{k,m}^{\text{prior}} \) is estimated, on each frame \( m \), as follows [4]:

\[
SNR_{k,m}^{\text{prior}} = \lambda_{1} \frac{\hat{S}_{k,m-1}}{E[N_{k,m}^2]} + (1-\lambda_{1}) \max \left( \frac{E[X_{k,m}^2]}{E[N_{k,m}^2]}, 1 \right) - 1.0 \tag{5}
\]

where

\[
E[X_{k,m}^2] = \lambda_{2} E[X_{k,m-1}^2] + (1-\lambda_{2}) E[X_{k,m}^2] \tag{6}
\]

2.2 Parameter estimation in the HSX speech coder

We consider the HSX speech coder at 2.4 kbits/s. This mixed excitation parametric coder was developed by Thomson-CSF in collaboration with the University of Sherbrooke [6]. It uses four parameters: the fundamental frequency (pitch), a voicing cut-off frequency between a low-pass voiced band and a high-pass unvoiced band, a set of linear prediction coefficients, and the energy.

The input signal is divided in 180-sample frames. The pitch and voicing parameters are estimated once per frame. Linear prediction parameters are estimated twice per frame by a 12th order Levinson-Durbin algorithm and converted to line spectral frequencies for more robust quantization. Energy is estimated four times per frame.

Pitch detection is computed from the normalized correlation of lower band (0-800Hz) semi-whitened input signal \( s'(j) \). Normalized correlation is defined by:

\[
C(i) = \frac{\sum_{j} s'(j)s'(j-i)}{\sqrt{\sum_{j} s'^2(j) \sum_{j} s'^2(j-i)}} \tag{7}
\]

where \( i \) represents lag allowed for pitch value:

\[
i_{\text{pitch}} = \arg \max_{i} C(i). \tag{8}
\]

Pitch decision is taken with one frame of look-ahead for robust pitch tracking, and is completed by a complex logical decision to avoid pitch doubling or halving.

For voicing decision, the frequency domain is separated in four subbands; in each one a voicing rate based on the normalized correlation is calculated. It quantizes the cut-off frequency on four available values, which defines the lower frequencies voicing domain and the higher frequencies unvoicing domain.

Energy is computed pitch-synchronously on four subframes and expressed in dB per sample. This parameter is not studied in this paper because of its slight contribution to the global intelligibility.

3. INFLUENCE OF BACKGROUND NOISE AND NOISE PRE-PROCESSING ON THE PERFORMANCE OF THE CODER

The exact parameters are derived from the clean speech signal. To measure objectively the influence of noise on the parameters estimation, we calculate the percentage of correct estimated pitch and voicing parameters, and the Log-spectral distance of auto-regressive coefficients for different segmental signal to noise ratios, \( SNR_{\text{seg}} \), with or without pre-processing. The ratio \( SNR_{\text{seg}} \) is
calculated on 180-sample frames only if speech is present. Three kinds of noise are considered: (i) white Gaussian noise, (ii) pink noise and (iii) babble noise. Speech proceeds from a French database used for intelligibility and quality tests. It consists in French words uttered by a male speaker and results are averaged on 15000 frames.

In order to measure, by means of a speech intelligibility test, the contribution of our speech enhancement system and each parameter to the global output intelligibility, we use a simplified rhyme test for French designed by the “Institut de Phonétique de l’Université d’Aix-Marseille” (IPAM) for Thomson-CSF. The intelligibility score it gives is statistically consistent with the standard comprehensive rhyme test for French language [7].

3.1 Pitch estimation

The pitch detection for noisy and pre-processed speech is considered to be correct when its value does not differ of more than 5% from the exact value and the result is averaged only on frames where speech is present.

3.2 Voicing estimation

Voicing is quantized only on four values so its detection for noisy and pre-processed speech is considered to be correct when its value does not differ of more than one position from the exact value, and the result is averaged only on frames where speech is present.

Figure 1 shows the percentage of correct pitch detection versus segmental signal to noise ratio. In the three cases, the noise reduction system improves the detection of pitch. This improvement goes up from 35% to 60% with white noise at -10dB. Lower improvements of pitch estimation are found with babble noise. Pitch estimation performance is very sensitive to the characteristics of additive background noise in the lower spectral band. Thus the pitch estimation is easier in the case of white noise.

3.3 Auto-regressive coefficients estimation

In the case of linear prediction coefficients, a number of methods exist to measure objectively the performance of the linear prediction. In our study, we use the Log-spectral distance [8], expressed in dB and averaged only on speech frames (Figure 3). This distance is an efficient measure of the spectral distortion that underlines any modification in the formants.
distance goes up to 9 dB with white noise at \(-10\) dB \(\text{SNR}_{\text{seg}}\). In white noise case, without noise pre-
processing, the estimation seems to be all the more disturbed as the noise spectral envelope is flat. The
modified Wiener filtering improves slightly the estimation when \(\text{SNR}_{\text{seg}}\) is below 10 dB. Using pink
noise, improvement occurs for very low \(\text{SNR}_{\text{seg}}\), below 0 dB. With babble noise, pre-processing decreases
slightly the estimation of auto-regressive coefficients whatever the noise level is. In a general manner, the
improvement brought by the noise reduction system is not significant for linear prediction coefficients.
 Comparable results are obtained using the Itakura-Saito distance.

3.4 Speech intelligibility
Speech intelligibility scores for noisy and noise reduction processed speech signals are presented in
Figure 4. At low \(\text{SNR}_{\text{seg}}\), global intelligibility of the output signal increases significantly when the WU
noise pre-processing is included, with white noise and pink noise. On the other hand, the intelligibility is
reduced by preprocessing with babble noise when \(\text{SNR}_{\text{seg}}\) decreases. The WU process seems to be not
suitable with babble noise whatever the noise level is.

![Figure 4. Score of intelligibility vs. \(\text{SNR}_{\text{seg}}\).](image)

In a previous paper we presented the contribution of each speech parameter to the global intelligibility [9].
We noted that all parameters contribute to the intelligibility but it seemed that the linear prediction
coefficients are the most important parameters. It is advisable to improve their estimation.
This study shows us that a noise pre-processing single-
sensor method is a good solution to improve the pitch and voicing estimation. On the other hand, linear
prediction coefficients are not improved by preprocessing, except for low \(\text{SNR}_{\text{seg}}\) with white and pink noises.

4. PROPOSED METHODS
Since the estimation of the linear prediction coefficients is essential in the speech intelligibility, we propose to
improve the noise reduction preprocessing to get a better estimation of these coefficients. We combine the
WU filter with an other filter, which takes into account the auto-regressive model of human voice. Our choice
fell on Lim’s filter based on the all-pole model of speech [10]. We combine linearly both filters as shown on
Figure 5. This global filter is performed on the spectral components of the observed signal obtained by
STDFT and it is expressed by:

\[ H_{k,m} = \alpha \cdot \frac{\text{SNR}_{\text{priori}}}{\text{SNR}_{\text{seg}}} \cdot \frac{1}{1 + \frac{\text{SNR}_{\text{seg}}}{\text{SNR}_{\text{seg}}} + (1 - \alpha)} \cdot \frac{\gamma_{k,m}^i}{\gamma_{k,m}^i + \gamma_{k,m}^s} \quad (9) \]

\(\gamma_{k,m}^i\) is the estimation of the power spectral density of the additive background noise, calculated during speech
pauses. \(\gamma_{k,m}^s\) is obtained by the following expression [10]:

\[ \gamma_{k,m}^s = \frac{\sum_{p=1}^{P} a_p e^{-2\pi x p T}}{1 - \sum_{p=1}^{P} a_p e^{-2\pi x p T}} \quad (10) \]

where \(T\) is the sampling period, \(a_p\) are the coefficients of the autoregressive model of order \(P\) (\(P=12\)), \(g\) is the
gain applied to the unit variance white noise applied on the model.

![Figure 5. Block diagram of the proposed method.](image)

In a first step \(\alpha\) is fixed to 0.5. The method is called the “Combined Pre-Processing” (CPP). Since Lim’s
filter performs better with additive white background noise, it is necessary either to pre-whiten the noise or to
take into account its spectral characteristics. In the second method we propose, \(\alpha\) depends on the noise
spectrum flattening. In 1979, Kay defined the spectral flattening factor \(\xi\) by the ratio between the geometric
average and the arithmetic average of the power spectral density to prove the flattening effect of
background white noise on the speech envelop [11]. The theory defines the flattening factor $\xi_n$, expressed in dB, at about 0 dB for a flat spectrum, and at $-60\, \text{dB}$ for a perfect tone [12, 13]. We calculate $\alpha$ as follows:

$$
\alpha = \begin{cases} 
\frac{1}{2} & \text{if } \xi_n \geq 0 \, \text{dB} \\
\frac{1}{2} \left(1 + \frac{\xi_n}{-30}\right) & \text{if } -30 \, \text{dB} < \xi_n < 0 \, \text{dB} \\
1 & \text{if } \xi_n \leq -30 \, \text{dB}
\end{cases}
$$

where $\xi_n$ is the noise flattening factor estimated during speech pauses. In this way, $\alpha$ is limited to 0.5 for flat noise spectrum to keep the benefits of the WU pre-processing, and it is equal to one for non uniform noise spectrum. This second method is called the “Modified Combined Pre-Processing” (MCPP).

5. EVALUATION RESULTS

5.1 Auto-regressive coefficients estimation

We present the results of the two methods proposed in section 4 on the estimation of the auto-regressive coefficients and the global intelligibility. Figure 6 presents the results for the CPP method and for noisy speech. The estimation of AR coefficients is slightly improved by the pre-processing with white noise and pink noise when $\text{SNR}_{seg}$ is respectively below 10 dB and 4 dB. The results of CPP method seem to be a bit better than the results of WU method. But the direct estimation from the noisy signal is better for babble noise, whatever the noise level is.

![Figure 6. Log-spectral distance vs. $\text{SNR}_{seg}$ with CPP method.](image)

Figure 7 shows the results for MCPP method. For white noise and pink noise the MCPP method improves the estimation of the AR coefficients for a $\text{SNR}_{seg}$ lower than 15 dB. Log-spectral distance decreases from 9 dB to 6.5 dB with white noise at $-10\, \text{dB}$, and from 7 dB to 5.6 dB with pink noise. Using babble noise and MCPP, the Log-spectral distance is nearly the same as using the noisy signal. So, the MCPP method seems to be more efficient for the AR estimation than the WU method and the CPP one.

![Figure 7. Log-spectral distance vs. $\text{SNR}_{seg}$ with MCPP method.](image)

5.2 Speech intelligibility

The improvement of the AR coefficients estimation is more suitable with the MCPP method. Let us see now the influence of this method on the global intelligibility. Figure 8 shows the intelligibility score with MCPP method. The presence of the pre-processing improves the intelligibility with all of the three noises. The intelligibility score goes from 82.5% to about 90.5% at $-10\, \text{dB}$ with white noise and from 87% to 92% with pink noise. With babble noise the intelligibility score of the processed signal is slightly higher than the intelligibility score of the noisy signal.

![Figure 8. Score of intelligibility vs. $\text{SNR}_{seg}$ with MCPP method.](image)
The improvement on the AR coefficients estimation brought by the MCPP method is confirmed by the improvement on the global intelligibility.

6. CONCLUSION

Single-sensor noise pre-processing methods are often considered to improve the performance of a low bit rate speech coder in the presence of acoustic background noise. We have shown that Wiener filtering under signal presence Uncertainty (WU) increases the intelligibility of coded speech. This improvement was confirmed by an objective analysis of the estimated speech coder parameters. However, it turned out that the improvement was limited in the case of the linear prediction coefficients.

In this paper, we have proposed two new noise pre-processing methods (called Combined Pre-Processing - CPP - and Modified CPP - MCPP). Those methods are based on a combination of two different classical noise pre-processing algorithms. We have shown that - unlike classical methods - MCPP - and to a lesser extent CPP - is not only capable of improving the pitch and voicing estimation of a parametric low bit rate speech coder, but also of improving the estimation of the linear prediction coefficients. This objective improvement in the parameters estimation is associated with an increase in intelligibility. In terms of computational complexity, since they are based on two classical single-sensor methods that share many processing blocks, both CPP and MCPP are fully compatible with a real time implementation on a PC or a current commercial DSP.

CPP and MCPP were optimized for the HSX speech coder. However, since all low bit rate speech coders are based on similar speech parameters, we believe that those methods could also apply to any other parametric speech coder.

We are currently beginning some work on developing a fully embedded speech enhancement processing. Such a procedure would not only take advantage of the processing already performed in the parametric coder, but also allow the coder to make use of the knowledge about the background noise when estimating the speech parameters.

7. REFERENCES


