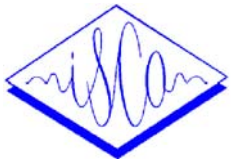

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

STUDY OF THE INFLUENCE OF NOISE PRE-PROCESSING ON THE PERFORMANCE OF A LOW BIT RATE PARAMETRIC SPEECH CODER

Gwénaël GUILMIN [1,2], *Régine LE BOUQUIN-JEANNÈS* [2], *Philippe GOURNAY* [1]

[1] Thomson-CSF Communications
66, rue du fossé blanc – B.P. 156
92231 Gennevilliers CEDEX – France
gwenael.guilmin@tcc.thomson-csf.com

[2] Laboratoire de Traitement du Signal et de l'Image
Université de Rennes 1
Bât. 22- Campus de Beaulieu
35042 Rennes CEDEX - France

ABSTRACT

This paper describes a prospective study of the contribution of a single-sensor noise pre-processing method, prior to coding, to the performance of a parametric low bit rate speech coder in adverse conditions. The 2.4kbits/s vocoder we use estimates four parameters: fundamental frequency, voicing, linear prediction coefficients and energy. Firstly, we study the influence of different noise levels on the estimated parameters with and without noise reduction system. Secondly, we measure the contribution of *(i)* each speech coder parameter and *(ii)* the speech enhancement system to the global output intelligibility. Finally, results show the interest of such a speech enhancement system for low bit rate parameter estimation and underline the interest to adapt different pre-processing techniques for each parameter estimation.

Keywords: noise reduction, parametric speech coding.

1. INTRODUCTION

For several years, particular attention has been 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 represent the most popular methods for single-sensor noise reduction [1][2][3]. Such systems are particularly useful in the field of low bit rate digital communications, since the speech parameters estimated by a parametric coder in the presence of background noise can be improved by a noise reduction system. Since the first well-known US and NATO standard LPC10, low bit rate speech coding has been greatly improved at the 2.4kbits/s data rate. New vocoders provide increased intelligibility and quality but they are still very sensitive to the acoustic speech environment and especially to additive background noise. The noise reduction procedure will probably be considered as a part of the next NATO low bit rate speech coding standard. So, it becomes clearly very attractive to fittingly combine noise processing and low bit rate speech coding. The 2.4kbits/s vocoder we use estimates four parameters: fundamental frequency, voicing defined by a cutoff frequency between a lower voiced band and an upper unvoiced band, linear prediction coefficients and energy. This parametric mixed-excitation 2.4kbits/s HSX coder was developed by Thomson-CSF in collaboration with the University of Sherbrooke [4].

Section 2 presents the noise reduction system we investigated, as well as the methods used by the coder to estimate the parameters. In section 3, we give results on some objective measures revealing the influence of different noise levels on the estimated parameters without noise reduction system. Using the same objective measures, we evaluate the improvement brought by noise pre-processing in the speech parameter estimation algorithms included in the vocoder. Section 4 describes the influence of the speech enhancement method on the intelligibility of coded speech. We mix "exact" parameters (*i.e.* derived from the original signal) and "noisy" parameters (*i.e.* derived from the noisy signal) at the input of the speech synthesis procedure of the decoder in order to measure, by means of a speech intelligibility test, their contribution to the global output intelligibility.

2. SPEECH ENHANCEMENT METHOD AND LOW BIT RATE SPEECH CODER

2.1. Wiener filtering under signal presence uncertainty

The noise reduction system we use corresponds to a modified Wiener filtering [5]. This algorithm is based on Ephraim and Malah estimator [6], given by the

minimization of the mean-square spectral error. Two hypotheses (H_1 and H_0) 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 amplitudes obtained by short-term discrete Fourier transform (STDFT), for each frame *m* of 180 points:

$$X_{k,m} = S_{k,m} + N_{k,m}. \quad (1)$$

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

$$H_{k,m} = W_{k,m}.G_{k,m} \quad (2)$$

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^{priori} and SNR^{post}). We can express the Wiener filter by:

$$W_{k,m} = \frac{SNR_{k,m}^{priori}}{SNR_{k,m}^{priori} + 1} \quad (3)$$

and the term of uncertainty by:

$$G_{k,m} = \frac{\Lambda_{k,m}}{\Lambda_{k,m} + 1} \quad (4)$$

where $\Lambda_{k,m}$ corresponds to the generalized likelihood ratio defined in [6] using the Gaussian statistical model assumed for the spectral components.

In our application the *a priori* SNR is estimated as follows [5]:

$$SNR_{k,m}^{priori} = \lambda_1 \frac{\hat{S}_{k,m-1}^2}{E[N_k^2]} + (1-\lambda_1)\max\left(\frac{E[X_k^2]_m}{E[N_k^2]} - 1; 0\right) \quad (5)$$

with:

$$E[X_k^2]_m = \lambda_2 E[X_k^2]_{m-1} + (1-\lambda_2) X_{k,m}^2 \quad (6)$$

where $E[X_k^2]_m$ represents the power spectral density of the noisy observation on the frame m. λ_1 and λ_2 are two forgetting factors equal to 0.98 and 0.3 respectively. $E[N_k^2]$ is an estimation of the noise power spectral density during silence periods using a first order estimator. $\hat{S}_{k,m-1}$ corresponds to the speech signal estimated in the previous frame $m-1$.

Such a system is often completed by a speech activity detection to update the estimation of the noise power spectral density. To do this detection, we use the uncertainty of speech signal presence calculated previously. The decision rule is based on the generalized likelihood ratio $\Lambda_{k,m}$ such as:

$$\overline{G}_m = \frac{2}{L-2}\sum_{k=1}^{\frac{L}{2}-1} G_{k,m} \underset{D_0}{\overset{D_1}{\gtrless}} \delta \quad (7)$$

where L and δ are respectively the discrete Fourier transform length ($L = 512$) and the level of speech detection ($\delta \approx 0.01$). D_0 and D_1 represent respectively the decision of absence and presence of speech for the frame *m*. In practice, the sampling rate is equal to 8kHz and the overlap between two successive frames is equal to 50%.

2.2. Estimation of speech parameters

The HSX speech coder uses a mixed harmonic and stochastic excitation [4]. 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:

$$C(i) = \frac{\sum_j s'(j).s'(j-i)}{\sqrt{\sum_j s'^2(j).\sum_j s'^2(j-i)}} \quad (8)$$

where i represents lag allowed for pitch value:

$$i_{pitch} = \arg\max_i C(i). \quad (9)$$

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 sub-frames and expressed in dB per sample. This parameter is not studied in this paper because of its little contribution to the global intelligibility.

3. INFLUENCE OF NOISE PRE-PROCESSING ON SPEECH PARAMETER ESTIMATION

To measure objectively the influence of noise on the parameters estimation, we calculate the percentage of corrected estimated pitch and voicing parameters, and the Log-spectral distance of auto-regressive coefficients for different segmental signal to noise ratios, SNR_{seg}, with or without pre-processing. These ratios are 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. In the following, the pre-processed case is denoted by WU. 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.

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.

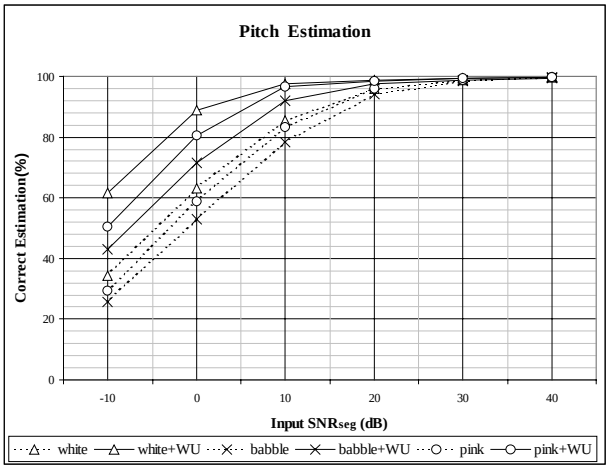

Figure 1. Correct pitch detection vs. SNR_{seg}.

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 provided 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.

3.2. 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 original value, and the result is averaged only on frames where speech is present.

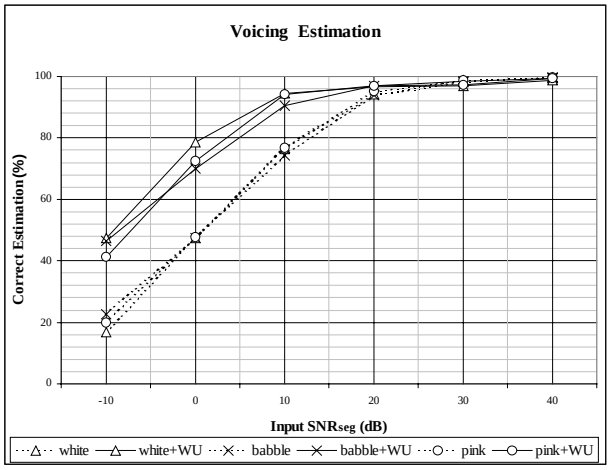

Figure 2. Correct voicing detection vs. SNR_{seg}.

Figure 2 shows the percentage of correct voicing detection versus segmental signal to noise ratio. Voicing detector is robust to all kinds of noise. Speech enhancement provides slightly the same improvement for white noise, pink noise and babble noise up to 25% at -10dB.

3.3. Auto-regressive coefficients estimation

In the case of linear prediction coefficients it exists many methods to measure objective performance of a linear prediction. In our study we use the Log-spectral distance [7]. If this distance gives a relevant measure of the spectral distortion, it is influenced by the formant positions, and averaged only on speech frames (Figure 3).

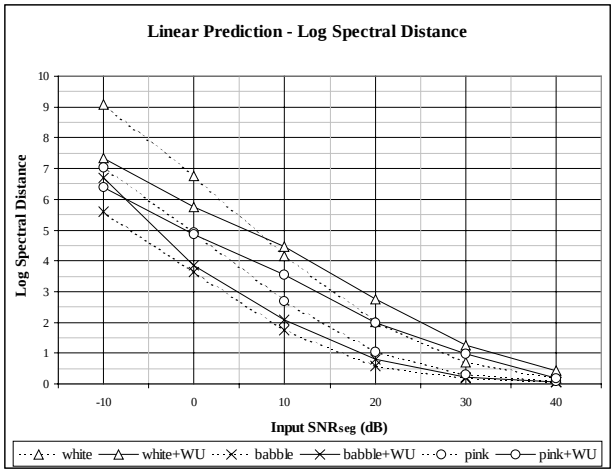

Figure 3. Log-spectral distance vs. SNR_{seg}.

Additive background noise decreases dramatically the estimation of the linear prediction coefficients especially for SNR_{seg} below 20dB. Log-spectral distance goes up to 9 with white noise at -10dB SNR_{seg}. So, without noise pre-processing, the estimation seems to be all the more disturbed because the noise spectral

envelope is flat. On the other hand, in white noise case, the modified Wiener filtering improves slightly estimation when SNR_{seg} is below 10dB. Using pink noise, improvement occurs for very low segmental SNR, below 0dB. 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.

4. SPEECH INTELLIGIBILITY

In order to measure, by means of a speech intelligibility test, the contribution of *(i)* our speech enhancement system and *(ii)* 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 classical, comprehensive rhyme test for French language.

The contribution of each speech parameter to the global intelligibility is estimated by mixing one "exact" parameter with the other "noisy" parameters at the input of the speech decoder. Let us indicate that pitch and voicing parameters are studied together because of their strong interaction in the speech coder.

Score	Original Speech	Noisy Speech	Exact AR	Exact Pitch-Voicing	Exact Energy
SNR_{seg}=10dB	97.26	94.70	95.45	95.13	94.85
SNR_{seg}=0dB	97.26	86.91	94.84	93.99	88.64

Table 1. Contribution of speech parameters on global intelligibility for pink noise.

Table 1 presents speech intelligibility scores, performed with 24 listeners, at 0dB and 10dB SNR_{seg}, with pink noise, when successively auto-regressive coefficients, pitch-voicing, and energy are extracted from original clean speech. All parameters contribute to the intelligibility but it seems that the linear prediction coefficients and the pitch-voicing are the most important parameters for speech intelligibility. This effect is more significant at low SNR_{seg}. The energy parameter appears to be less important whatever the SNR_{seg} is.

Score	Original Speech	Noisy Speech	Pre-Processing WU
SNR_{seg}=10dB	97.26	94.70	97.14
SNR_{seg}=0dB	97.26	86.91	92.66

Table 2. Intelligibility of noisy and pre-processed speech signals for pink noise.

Speech intelligibility scores for pink noisy and noise reduction processed speech signals are presented in Table 2. Global intelligibility of the output signal increases significantly when the WU noise pre-processing is included.

5. CONCLUSION

All parameters contribute to the speech quality and intelligibility but we find that the linear prediction coefficients and the pitch-voicing are the most important parameters for speech intelligibility. Noise pre-processing single-sensor method is indeed a good solution to improve the performance of a low bit rate speech coder in the presence of acoustic background noise. According to the informal listening tests we conducted, it comes out that the pre-processing system enhances speech quality. This study confirms the improvement in the speech coder parameter estimation, even if this improvement is objectively limited in the case of the linear prediction coefficients. On the other hand, it increases the global intelligibility of coded speech.

The speech enhancement system we tested is suitable for pitch and voicing estimation but another process could be used for the linear prediction coefficients. The next step of our work would be to test one distinct speech enhancement processing for each parameter computed in the speech coder. The global system must be low complexity and low delay for real-time applications.

6. REFERENCES

[1] BEROUTI M., SCHWARTZ R., and MAKHOUL J., *Enhancement of Speech Corrupted by Acoustic Noise*, ICASSP, pp. 208-211, 2-4 April 1979.

[2] BOLL S.F., *Suppression of Acoustic Noise in Speech Using Spectral Subtraction*, IEEE Trans. on ASSP, vol. ASSP-27, n°2, pp. 113-120, April 1979.

[3] LIM J.L., and OPPENHEIM A.V., *Enhancement and Bandwidth Compression of Noisy Speech*, Proceedings of the IEEE, vol. 67, n°12, pp. 1586-1604, December 1979.

[4] LAFLAMME C, SALAMI R., MATMTI R., and ADOUL J-P., *Harmonic-Stochastic Excitation (HSX) Speech Coding below 4kbits/s*, ICASSP, pp. 204-207, 1996.

[5] AKBARI AZIRANI A., LE BOUQUIN R., and FAUCON G., *Speech Enhancement Using a Wiener Filtering under Signal Presence Uncertainty*, EUSIPCO, pp. 971-974, 1996.

[6] EPHRAIM Y., and MALAH D., *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator*, IEEE Trans. on ASSP, vol. ASSP-32, n°6, pp. 1109-1121, December 1984.

[7] BASSEVILLE M., *Distance Measures for Signal Processing and Pattern Recognition*, Signal Processing 18, pp. 349-369, 1989.