A Simple Algorithm for the Restoration of Clipped Speech Signal

Abdelhakim Dahimene, Mohamed Noureddine and Aarab Azrar
Electrical and Electronic Engineering department, Boumerdes University
Boumerdes, Algeria, 35000
E-mail: dahimenehakim@yahoo.fr

Keywords: speech signal, clipped speech, restoration, interpolation, linear prediction, least square method, Kalman filter

Received: August 6, 2007

This paper deals with the problem of peak clipped speech. Our basic assumption is that the clipped speech is voiced and can be linearly predicted with a high accuracy. The coefficients of linear prediction are computed using two different algorithms: a least square direct method and a recursive Kalman filter. The speech reconstruction is accomplished using backward prediction.

Povzetek: Predstavljen je algoritem za obnavljanje zvočnega signala.

1 Introduction

Speech acquired by personal computer sound cards is often confronted with two main problems: DC level wandering and peak clipping. While building a data base for our speech recognition project, we have been confronted with both problems. The first one is easily eliminated by simple linear processing but the second one requires more complex algorithms. Peak clipping is fundamentally a non linear distortion. It is characterized by the fact that several successive values of the signal disappear and are replaced by a constant. However, it happens that speech signal is highly predictable. So, in essence, peak clipped speech restoration is a problem of interpolation since we are trying to find missing values by using the properties of the signal itself. There exist several methods of interpolation: polynomial (Lagrange, Newton), spline, etc. In the case of peaked clipped speech, an appropriate method is statistical interpolation [1].

2 Justification of the method

When there is no a priori information on the signal, the classical numerical interpolation methods (polynomial and spline) should be used. Band limited interpolation [2] uses only the fact that the signal is band limited. Statistical interpolation based on linear prediction [2, 4] uses the fact that that speech signal is highly predictable. A speech segment is composed of a sequence of voiced, unvoiced and silence (noise) segments [2]. The type of speech signal that has the greatest probability for being peak clipped is voiced speech [2, 3]. Figure 1 represents a scatter plot of voiced, unvoiced and silence mean magnitude and zero crossing rate of segments of speech. Voiced speech segments are indicated by the letter "V", unvoiced segments by the letter "U" and the silent segments by the letter "S". It shows clearly that the voiced signals cluster at high mean magnitude values.

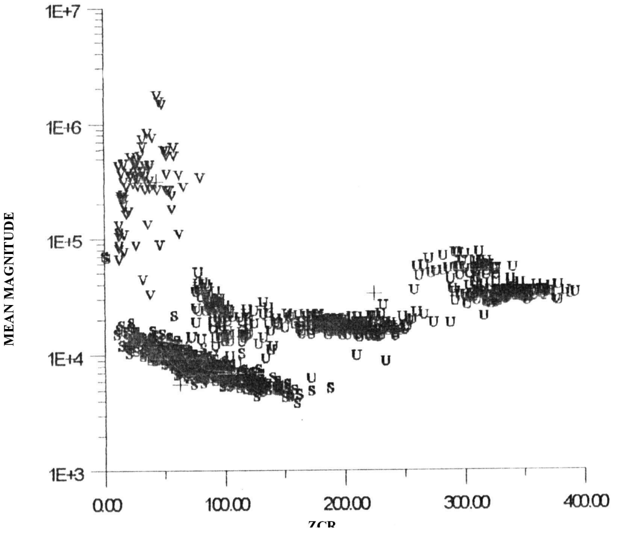

Figure 1: Mean magnitude and ZCR scatter plot [3]

The mean magnitude is defined as:

$$M_n = \sum_{m=-\infty}^{\infty} |x(m)|\, w(n-m) \tag{1}$$

where $w(n)$ is a rectangular window of length 256 samples and the zero crossing rate (ZCR) is:

$$ZCR(n) = \sum_{m=-\infty}^{+\infty} \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m) \tag{2}$$

Fortunately, voiced speech happens to be quite predictable. Voiced speech follows quite closely the linear prediction equations [4, 5]. Commercial software like DC-6, from Diamond Cut products, use low order linear prediction for clipped audio signal restoration and the problem of audio signal interpolation have also been addressed by Vaseghi [1] who uses linear prediction from adjacent samples and samples one period away (audio signals are assumed to be periodic).

Voiced speech can be considered as a quasi periodic signal. It can be modelled as the output of a linear time invariant system (during few milliseconds, the system can safely be assumed to be time invariant) driven by a periodic train of impulses. In this case, a quite general formulation of the signal will be:

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} + \sum_{k=0}^{p} b_k u_{n-k} \tag{3}$$

where the signal u_k is equal to 1 every T seconds and zero otherwise, T being the pitch period. a_k and b_k are respectively the recursive and the non recursive parameters of the above production filter of order p. So, within a pitch period (N_T samples) and after p samples, we can write:

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} \tag{4}$$

The above equation breaks down in the part of the speech signal that is clipped. So, if we start the time axis at the beginning of a pitch period and if we call N_T the number of samples within the pitch period, we can write:

$$x_n = \sum_{k=1}^{p} a_k x_{n-k} \quad ; \quad p \le n \le N_T \qquad for \ |x_n| < X_{\max} \tag{5}$$

$X_{\max}$ being the saturation value.

3 The proposed restoration algorithm

The proposed algorithm for clipped speech restoration is going to be based on linearly predicting the missing values using equation (4). So, the algorithm consists of two following steps:

- Computation of the prediction coefficients a_k .
- Linear prediction of the missing values.

3.1 Computation of the prediction coefficients

The computation of the prediction coefficients a_k can be accomplished either by using a least square solution or by using a recursive algorithm based on Kalman filtering.

For the least square algorithm, we can use equation (5) and build the following matrix vector equation relating speech samples x_k:

$$\begin{pmatrix} x_{p+1} \\ x_{p+2} \\ . \\ . \\ x_{N_T} \end{pmatrix} = \begin{pmatrix} x_p & x_{p-1} & . & . & x_1 \\ x_{p+1} & x_p & . & . & x_2 \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{N_T-1} & x_{N_T-2} & . & . & x_{N_T-p} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ . \\ . \\ a_p \end{pmatrix} \tag{6}$$

in which all the rows such that $|x_k| = X_{\max}$ are deleted. Equation (6) can be written as:

$$\mathbf{b} = \mathbf{X}\ \mathbf{a} \tag{7}$$

and the least square solution of equation (7) can be obtained as [6]:

$$\mathbf{a} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{b} \tag{8}$$

Another approach to the evaluation of the prediction coefficient is the following sequential algorithm (Kalman filter) [7, 8] based on the subsequent set of equations and on an autoregressive model. Consider the next state equation:

$$\mathbf{a}(k+1) = \mathbf{a}(k) + \mathbf{w}(k) \tag{9}$$

where $\mathbf{a}(k) = \left(a_1, a_2, \ldots, a_p\right)^T$ and $\mathbf{w}(k)$ is a white stationary sequence.

The observation model is:

$$x_n = \sum_{i=1}^{p} a_i\, x_{n-i} + b_0\, u_n \tag{10}$$

and let us consider $\mathbf{C}(k-1) = \left[x_{k-1}, x_{k-2}, \cdots, x_{k-p}\right]$

then the observation model becomes:

$$z(k) = x_k = \mathbf{C}(k-1)\mathbf{a}(k) + v(k) \tag{11}$$

if it is taken that: $v(k) = b_0 u_k$

then, starting from an initial estimate $\hat{\mathbf{a}}(0)$, we obtain the following recursion:

$$\hat{\mathbf{a}}(k+1) = \hat{\mathbf{a}}(k) + \mathbf{K}(k+1)[x_{k+1} - \mathbf{C}(k)\hat{\mathbf{a}}(k)] \tag{12}$$

where $\mathbf{K}$ is the Kalman gain given by:

$$\mathbf{K}(k+1) = \frac{\mathbf{V}_{\hat{\mathbf{a}}}(k)\mathbf{C}^{\mathbf{T}}(k)}{b_0^2 + \mathbf{C}(k)\mathbf{V}_{\hat{\mathbf{a}}}(k)\mathbf{C}^{\mathbf{T}}(k)} \tag{13}$$

and the matrix $\mathbf{V}_{\hat{\mathbf{a}}}$ is the variance matrix of the estimator $\hat{\mathbf{a}}$ and is given by the following equation:

$$\mathbf{V}_{\hat{\mathbf{a}}}(k+1) = [\mathbf{I} - \mathbf{K}(k+1)\mathbf{C}(k)]\mathbf{V}_{\hat{\mathbf{a}}}(k) + \mathbf{V}_{\mathbf{w}} \tag{14}$$

where $\mathbf{V}_{\mathbf{w}}$ is the variance matrix of the white noise process $\mathbf{w}(k)$.

The algorithm can be initialized by: $\mathbf{V}_{\mathbf{w}} = \sigma^2\mathbf{I}$, $\mathbf{V}_{\hat{\mathbf{a}}}(0) = \mathbf{0}$ and $b_0 = 1$ and stopped by using the criterion:

$$\left\|\hat{\mathbf{a}}(k+1) - \hat{\mathbf{a}}(k)\right\|^2 \le \varepsilon \tag{15}$$

The stopping criterion can also be used for pitch detection because it is evident that the above norm will be large while being in a clipped part, since the autoregressive model will not be valid.

3.2 Interpolation of the missing samples

For the computation of the missing samples, equation (4) can be used starting from p previous samples. This interpolation is referred to as forward. The missing samples can also be predicted from p samples that follow the missing part. The first sample can be obtained by solving equation (4) as:

$$x_{n-p} = \sum_{i=1}^{p} \alpha_i x_{n-p+i} \qquad ; \quad p+1 \le n \le N_T \tag{16}$$

where the coefficients α_i are computed from the coefficients a_i using:

$$\alpha_1 = -\frac{a_{p-1}}{a_p} \ ; \ \cdots \ ; \ \alpha_{p-1} = -\frac{a_1}{a_p} \ ; \ \alpha_p = \frac{1}{a_p} \tag{17}$$

Consequently, the reconstruction is done using backward interpolation.

4 Results

In order to test the previously defined algorithms, we are going to use synthetic and natural speech. The natural speech comes from a very large database of speech samples that were collected for the construction of a speech recognition system in colloquial Algerian Arabic [9]. The pitch frequency is about 100 Hz for male speaker and about 220 Hz for a female one. This corresponds to a pitch period T being between 4.5 ms to 10 ms. So, if a reliable estimation of the prediction parameters is desired, we need a fairly high sampling frequency. For example, a sampling frequency of 10 kHz (sampling period of 100 µs) will provide between 45 and 100 samples for a pitch period. A sampling frequency of 44.1 kHz (sampling period of 22.73 µs) is chosen, which provides between 198 and 440 samples for a pitch period, which is quite reasonable. Also, in all of the following tests, the speech signal is normalized to a maximum value of one.

4.1 Synthetic speech

First the algorithm is tested with a synthetic vowel. The choice of synthetic speech is motivated by the fact that it follows exactly the linear prediction model. The vowel /a/ is generated using the following formants [5]:

- The frequencies (F_i) and the bandwidths (BW_i) necessary to specify each formant are shown in the following table.

Formant	$F_i(Hz)$	$BW_i(Hz)$
1	730	60
2	1090	100
3	2440	120
4	3500	175
5	4500	281

Table 1: Formant Frequencies and Bandwidths [5]

- The pitch frequency is 120 Hz (male speaker), which corresponds to $N_T = 367$ samples. Figure 2 shows few periods of the synthetic vowel /a/. The prediction order is set to $p = 10$.

This signal is clipped to a level of ±0.5 and restored using both methods (least square and Kalman filter method).

Figure 3 shows one pitch period of the clipped signal. A window of at least 75 samples following the clipped region is used to compute the predictor coefficients.

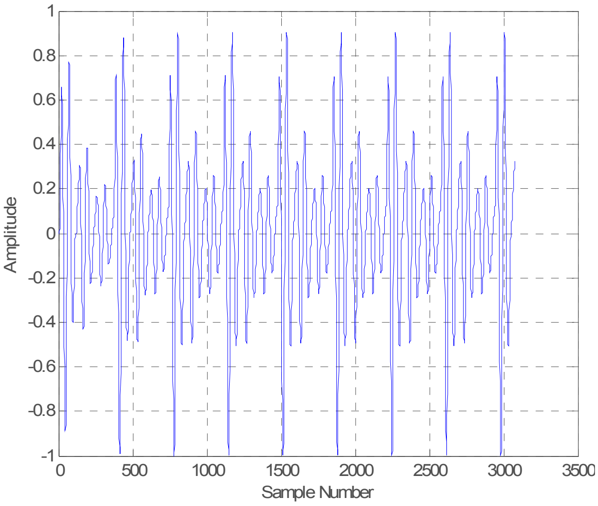

Figure 2 : Synthetic Vowel /a/ Normalized Amplitude Waveform

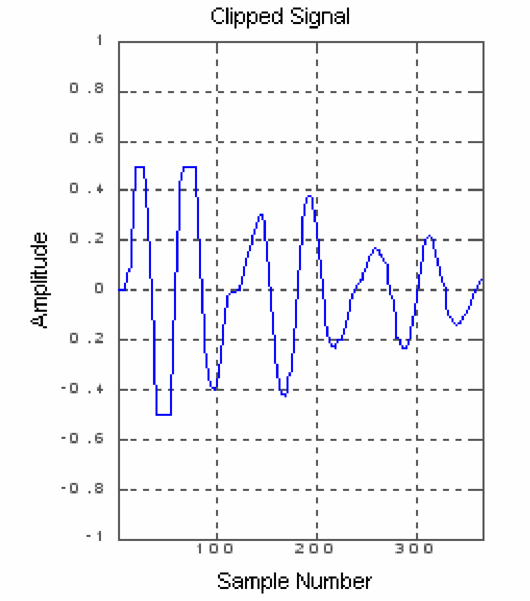

Figure 3 : Clipped artificial vowel

The least square computation of the prediction coefficients along with both forward and backward reconstruction produces the error plots shown in figure 4. It can be seen that the backward error is much smaller than the forward one. Also, the error occurs at the end of the reconstruction. The error can be reduced by performing both reconstructions and averaging the results. However, since the error is essentially a high frequency signal, simple low pass filtering after backward reconstruction yields the same result.

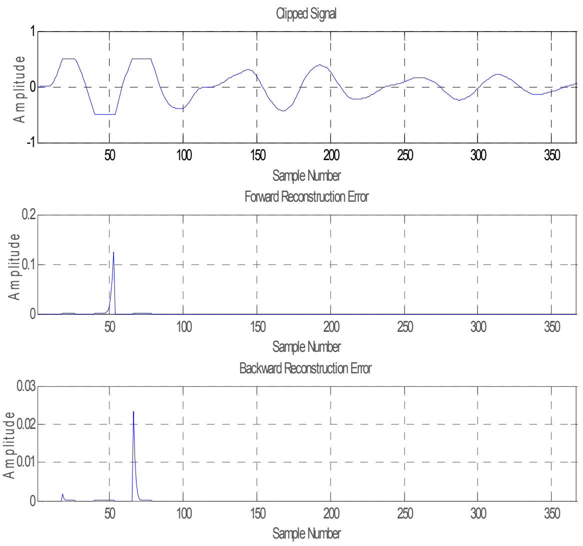

Figure 4 : Forward and Backward Reconstruction Error for Synthetic Speech

Kalman filter is also used for the estimation of the prediction parameters. The stopping criterion of the recursive Kalman algorithm is defined as: $\|a(k+1)-a(k)\|^2 < \varepsilon$, where ε is a small positive number that describes the convergence of the algorithm. From the plot of $\|a(k+1)-a(k)\|^2$ over 04 pitch periods of the signal (under: $\sigma = 0.1$ *and* $b_0 = 1$) shown in figure 5, it appears that the value ε=0.00008 is acceptable.

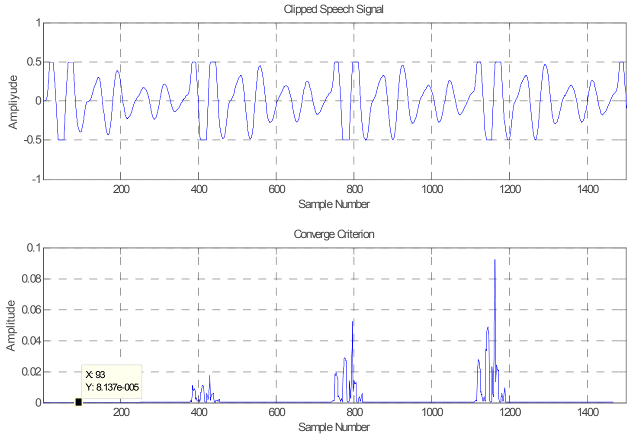

Figure 5 : Kalman Filter Convergence Criterion Plot

As stated before, the above convergence criterion ($\|a(k+1)-a(k)\|^2$) can be used for pitch detection. This is well illustrated in figure 5, the large values occurs at the clipped parts generally located at the beginning of the pitch period. After several tests, the following initial values: $\sigma = 0.1$, $b_0 = 1$, $\hat{\mathbf{a}}(0) = \mathbf{0}$ have been selected. After estimation of the prediction parameters and backward prediction, the error is drawn in figure 6.

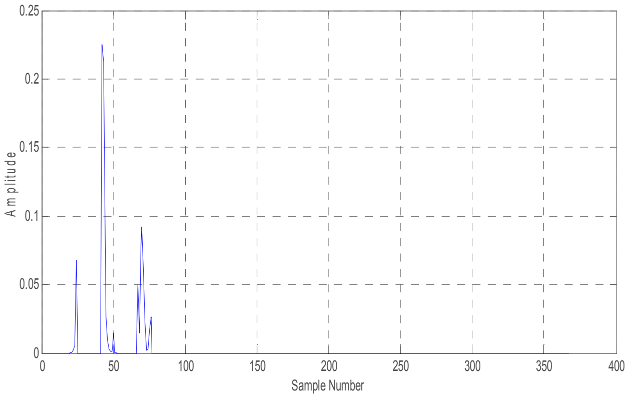

Figure 6 : Kalman Filter Error signal waveform

We observe that the error using the Kalman filter estimation is much larger that the one using the least square method. Another problem that is encountered is the large computation time. So, in the following tests, the results obtained by the least square method are the only ones that will be presented.

4.2 Artificially clipped natural speech

After being applied on a synthetic speech, the proposed technique of interpolation (least square evaluation of the parameters and backward reconstruction) is applied on a voluntarily clipped natural speech. The unclipped signal is taken as a reference when evaluating the reconstruction process precision.

The used recorded speech signal consists on numbers pronounced in Algerian Arabic, sampled at 16 KHz, taken from the database [9]. An audio processing software (Cool Edit Pro 2.1 from Syntrillium Software Corporation) is used to adjust the sampling frequency to 44.1 KHz.

Since speech is a time varying signal (a concatenation of different sounds with different characteristics) and in order to have a good estimation of the prediction parameters, the following method based on the detection of clipped samples is used: after each detection of a clipped sample, an adjacent segment of enough number of successive unclipped samples (ex.: in our case 75 samples) is considered. If this condition is satisfied the reconstruction process that uses the least square algorithm for the estimation of the prediction parameters will be applied. Otherwise, the procedure is repeated. Figure 7 shows the different steps of signal processing. The original speech and the reconstructed one are practically identical. Figure 8 shows the reconstruction error for the natural speech where it can be observed that the error is a high frequency signal with a small peak magnitude. So, a simple low pass filter will eliminate completely the error.

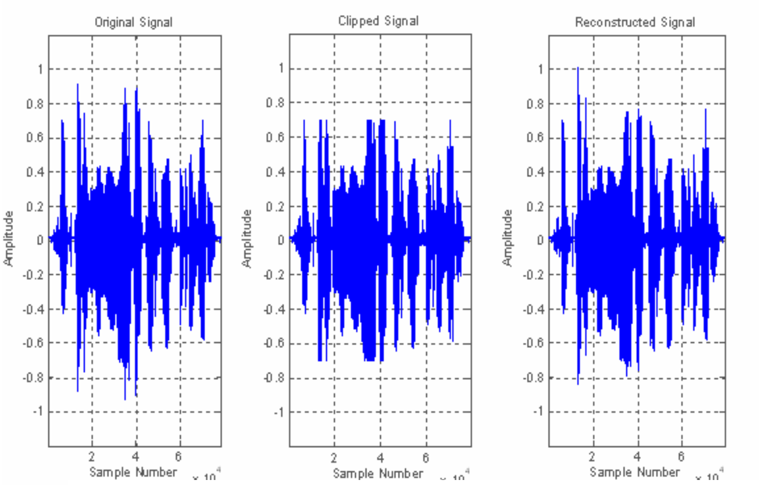

Figure 7 : Artificially Clipped Natural Speech Reconstruction

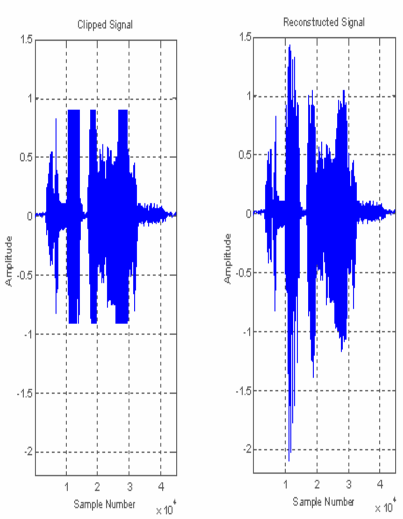

Figure 9 : Clipped Natural Speech Backward Reconstruction using Least Square Estimation

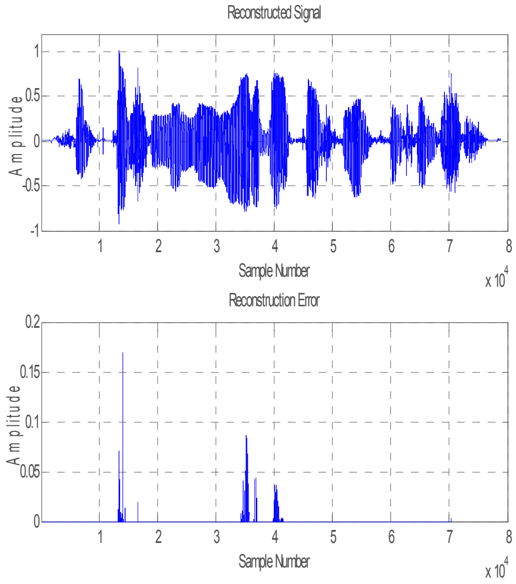

Figure 8 : Artificially Clipped Natural Speech Reconstruction Error

4.3 Clipped natural speech

The final test is performed on clipped natural speech. Figures 9 and 10 show the clipped and the reconstructed signal. It is impossible in this case to present an error plot due to the absence of the original unclipped signal. The only comment that can be made about the above plots is that the reconstructed signal looks like an unclipped signal. Since there is no reference to objectively evaluate the performance of the algorithm, a subjective criterion is used for judging the quality of the restoration. The speech samples (clipped and restored) were presented to several listeners and they were asked to evaluate the intelligibility of the message by giving a grade between zero and five (zero meaning completely unintelligible and five meaning very clear). The result is a great improvement in intelligibility. The clipped signal was given an average grade of about two while the restored signal received a grade that varied between four and five.

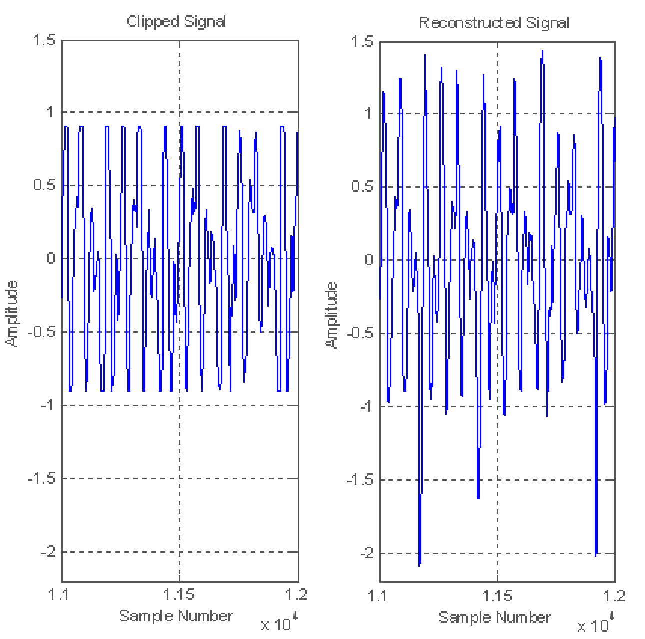

Figure 10 : Zoomed segment of the reconstruction process.

4.4 Discussion

It is quite hard to provide a figure of merit for the method other than the plot of the error signal between the unclipped and the reconstructed speech signal. We can observe from the plots in figure 4 and 8 that two parameters determine the quality of the reconstruction: the amplitude and the duration of the error spikes. We can resume both parameters in the following quality factor:

$$\Gamma = \frac{1}{N_c}\sum_{k=1}^{N_c}|e(k)| \qquad (18)$$

where $e(k)$ is the error signal between the unclipped and the reconstructed speech signal and N_c is the number of clipped samples. Figure 11 shows the quality factor for forward and backward reconstruction as a function of the number of clipped samples per pitch period for synthetic speech. For clipped natural speech, it of course impossible to provide such data. For artificially clipped natural speech, the quality factor curve using backward reconstruction is so close to the one for forward reconstruction for synthetic speech that it is impossible to provide a separate plot. From the different plots (figure 4, 8 and 11), we can conclude that the estimation of parameters using least square followed by a backward reconstruction offers the best results in term of accuracy.

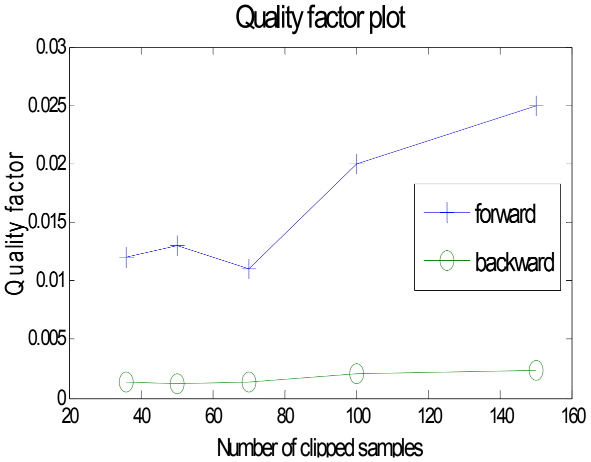

Figure 11 : Plot of quality factor vs. number of clipped samples for synthetic speech.

5 Conclusion

In this paper, an algorithm for clipped speech restoration using linear prediction has been presented and tested. It is able to restore completely the clipped speech. Two different methods for estimating the prediction parameters have been tested. The first one consists on block least square estimation while the second one is a recursive method. It appears that the recursive method is pitch synchronous but quite inefficient while the block least square is very efficient and very precise. The block least square method followed by backward prediction has been implemented as part of a larger program for speech pre-processing in view of recognition and the results are a great improvement in the recognition rate.

6 References

[1] Vaseghi, S.V., “Advanced Signal Processing and Digital Noise Reduction*”, John Wiley and Teubneur* 1996.

[2] Rabiner, L. R., Shafer, R. W., “Digital Processing of Speech Signals”, Prentice Hall, New Jersey, 1978.

[3] Brakta, N. and Hadibi, M., “Pattern Recognition Techniques Applied to Speech Recognition”, Final Year Project, INELEC, 1999.

[4] Makhoul, J., "Linear Prediction, A Tutorial Review", Proceeding of the IEEE, (63), pp. 561-580, April 1976.

[5] Chandra, S. and Lin, W. C., "Experimental Comparison between Stationary and Non Stationary Formulations of Linear Prediction Applied to Voiced Speech Analysis", IEEE trans. of ASSP, (22)6, Dec. 1974.

[6] Lawson, C.L., and Hanson R.J., “Solving Least Squares Problems”, Prentice Hall, 1974.

[7] Srinath, M.,D. and Rajasekaran, P., K., “An Introduction to Statistical Signal Processing with Applications”, John Wiley & Sons, 1979.

[8] Gueguen, C., J., and Carayannis, G., "Analyse de la Parole par Filtrage Optimal de Kalman", Automatisme, Tome XVIII, March 1973, pp.99-105.

[9] Reggab, M., “Continuous Speech Recognition Using Hidden Markov Models, Application to Colloquial Algerian Arabic*”*, Unpublished Magister Thesis, DGEE, FSI, Université M'Hamed Bouguerra, Boumerdes, Algeria, June 2004.