MELECON '85 / Volume II: Digital Signal Processing
A. Luque, A.R. Figueiras Vidal, V. Cappellini (eds.)
Elsevier Science Publishers B.V. (North-Holland) / © IEEE 1985

# ON CRITICAL ASPECTS OF PARAMETERS EVALUATION IN MULTI-PULSE LPC SPEECH SYSTEMS

M.G. Di Benedetto*, P. Mandarini*, R. Viola**

(*) Infocom Department, University of Rome
Via Eudossiana, 18 - 00184 Rome (Italy)
(**) ITALSPAZIO, Consorzio Industriale per le Attivita'
Spaziali, Via V.E. Orlando, 83 - 00185 Rome (Italy)

## ABSTRACT

Speech synthetizers, based on "multipulse excitation" are characterized by a good voice reproduction quality.
Previous work have suggested sub-optimal iterative procedures addressed to identify the positions and amplitudes of the excitation pulses in an environment substantially stationary.
In this paper the "unsolved ambiguities" related to the non-stationary nature of the speech signal are shown.
Then a new approach is suggested, providing the distance between the original signal and the synthetic one, that allows an improvement of the signal to coding noise ratio and reduces the computation time.

## 1. INTRODUCTION

The speech coding and synthesis through a prediction analysis introduces a broad class of synthetizers, likely represented in the general scheme of Figure 1a.

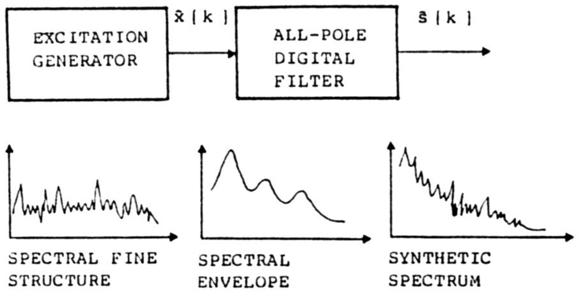

FIGURE 1a - GENERAL LPC SYNTHETIZER

The all poles filter reproduces the short time speech spectral envelope and the filter coefficients are obtained through linear prediction of speech signal, that leads to the well-known covariance or autocorrelation procedures [1], [2].
The excitation generator strongly influences the speech quality. In the traditional LPC vocoders, the excitation is constituted, according to the nature of the incoming speech signal, either by pulse generator for voiced sounds, or by a white noise generator for unvoiced sounds (Figure 1b).
A multipulse excitation has been proposed by Atal and Remde [3], to improve the voice quality. This excitation is constituted by a stream of pulses in opportune positions and with suitable amplitudes. Obviously a more detailed source description implies an increase of the bit rate from 2.4 Kbps (traditional LPC vocoder) to 9.6 - 12 Kbps.

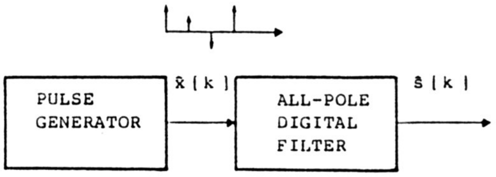

FIGURE 1b - MULTIPULSE SYNTHETIZER

The optimal amplitudes and positions can be theoretically calculated by an "analysis through synthesis" procedure which minimizes a frequency weighted mean square error between the original speech waveform and the synthetic one (Figure 2). The error weighting $\mathcal{W}\{\cdot\}$ takes into account the auditory masking properties of the human ear.
The error minimization procedure leads to a non linear system. In order to solve this problem a sub-optimal iterative procedure is proposed by Atal and Remde [3]. This procedure performes the error minimization searching a pulse (amplitude and position) at a time, with no further recalculation of the amplitudes and positions of the previously selected pulses.

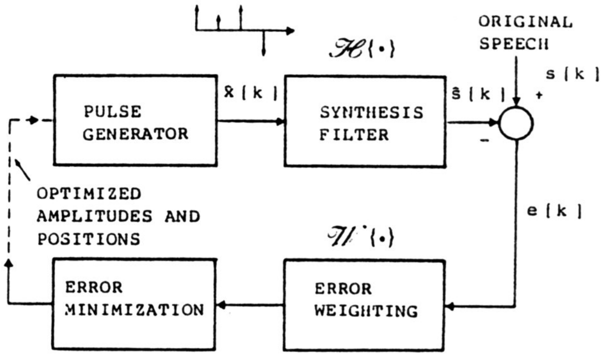

FIGURE 2 - ANALYSIS THROUGH SYNTHESIS PROCEDURE FOR OPTIMAL PULSE SEARCH

CH 2185-7/85/0000-0115(B)/$01.00 © 1985 IEEE

In practical situations, taking into account the non-stationary nature of the speech signal, different error criterium, i.e. different distance measures between the original and the synthetic signals, can be derived from the previously mentioned approach.
In this paper a new distance measure is proposed. This distance tries to overcome the problems connected with a non-stationary speech model, improves the synthetic speech quality and reduces the calculation amount.

## 2. ERROR CRITERIA FOR A SUB-OPTIMAL PULSE SEARCH

Starting from the general error criterium outlined in introduction, a non-ambiguous distance can be derived assuming that:

i) The transformation $\mathscr{W}\{\cdot\}$ is time-invariant.

ii) The speech signal is time limited, say k=0 to N.

In this hypothesis, $\mathscr{W}\{\cdot\}$ can be described by the transfer function:

$$H(f) = \frac{1}{1-\sum_{m=1}^{M} z^{-m}\,p[m]}\Bigg|_{z=e^{-j2\pi f/2W}} \qquad (1)$$

The weighting function is generally described, [3], [5], by means of a linear filter $\mathscr{W}\{\cdot\}$ , time-invariant in this case, whose transfer function is:

$$W(f) = \frac{1-\sum_{m=1}^{M} z^{-m}\,p[m]}{1-\sum_{m=1}^{M} z^{-m}\,\gamma^{m}\,p[m]}\Bigg|_{z=e^{-j2\pi f/2W}} \qquad (2)$$

in which $\gamma$ is a constant in the range from 0.8 to 0.9. The distance to be minimized by the multipulse sequence x[k] is given by:

$$D=\frac{1}{2W}\int_{-W}^{W}\Big|E(f)\,W(f)\Big|^{2}df = \sum_{k=0}^{\infty}\Big\}e[k]*w[k]\Big\{^{2} \qquad (3)$$

where E(f) = S(f)-Ŝ(f) and S(f), Ŝ(f) are the DFT of S[k], Ŝ[k] respectively.
Unlikely, in real situations, the speech signal has an indefinite length and the prediction coefficients are updated every N samples; therefore the transformation $\mathscr{H}\{\cdot\}$ and $\mathscr{W}\{\cdot\}$ can be considered to be stationary only in limited time intervals, and the environment becomes non-stationary.
Nevertheless a non-stationary approach can be avoided if the time-invariance of the prediction coefficients in each frame (N samples length) is taken into account. With this assumption two possible approaches have been presented [3], [4].

The first approach leads to choice the multipulse sequence $\hat{x}^{[j]}[k]$ (*) for each frame j, which minimizes in frequency domain the distance (3) evaluated for frame j.
In formula $x^{[j]}[k]$, is chosen in order to minimize:

$$D_1^{(J)}=\frac{1}{2W}\int_{-W}^{W}\Big|E^{(J)}(f)\,W^{(J)}(f)\Big|^{2}df=\sum_{k=0}^{\infty}\Big\}e^{[J]}[k]*w^{(J)}[k]\Big\{ \qquad (4)$$

where the simbol * stands for convolution in time and $W^{(j)}[k]$ is the impulse response of the weighting filter.
In a second approach, [4], $\hat{x}^{[j]}[k]$ is selected for each frame j by minimizing the weighted error along the frame duration. In other words, $\hat{x}^{[j]}[k]$ is chosen in order to minimize the quantity:

$$D_2^{(J)}=\sum_{k=0}^{N-1}\Big\}e^{[J]}[k]*w^{(J)}[k]\Big\{^{2} \qquad (5)$$

Both methods (4), (5) are not satisfactory from a theoretical point of view. To focuse this aspects let us introduce the prediction residual, [1]:

$$x(k)=\mathscr{H}^{-1}\Big\}s[k]\Big\{ \qquad (6)$$

and then let decompose s[k] and e[k] as follows (Figure 3):

$$s[k]=\sum_{J=1}^{\infty}s^{(J)}[k] \text{ with: } s^{(J)}[k]\doteq\mathscr{H}\Big\}\hat{x}^{[J]}[k]\Big\{ \qquad (7)$$

$$e[k]=\sum_{J=1}^{\infty}e^{(J)}[k] \text{ with: } e^{(J)}[k]\doteq s^{(j)}[k]-\mathscr{H}\Big\}\hat{x}^{(J)}[k]\Big\{ \qquad (8)$$

After, let us suppose that the lenght of impulse response of all filters used is lower than N so that the duration of signal (Figure 3):

$$\delta^{(J)}[k]=\begin{cases}0 & \text{for } k<N\\ e^{(J)}[k] & \text{for } k\geq N\end{cases} \qquad (9)$$

is lower than N, also in case of perceptual weighting. With the previous notations and hypotheses the (4), (5) become respectively:

(*) In that follows, we indicate by j the current frame index, by a superscript (j) a generic quantity related with the frame j, by superscript [j] a generic signal multiplied with a rectangular window having the same frame length. In addition, the subscript j represents signal filtered by the weighting function (2), (with the parameters calculated in the frame j), and the index k ranging from 0 to N-1 in the frame j.

$$D_1^{(j)} = \sum_{K=0}^{\infty} \left\{ e_j^{(j)}[k]^2 + \delta_j^{(j-1)}[k]^2 + +2\delta_j^{(j-1)}[k] \cdot e_j^{(j)}[k] - \delta_j^{(k)}[k]^2 \right\} \tag{10}$$

$$D_2^{(j)} = \sum_{K=0}^{N-1} \left\{ r_j^{(j)}[k] - \sum_{l=1}^{L} a_l h_j[k-m_l] \right\} \tag{11}$$

where:

$$r_j^{(j)}[K] = \left\{ s^{(j)}[k] + \delta^{(j)}[k] \right\} * W^{(j)}[k] = \delta_j^{(j)}[k] + \delta_j^{(j-1)}[k] \qquad 0 \leqslant k < N \tag{12}$$

$$h_j^{(j)}[k] = \mathscr{W}^{\cdot} \left\{ \mathscr{H} \left\{ u_o[k] \right\} \right\} \tag{13}$$

and $a_e$ and $m_e$ are respectively the pulses amplitude and positions in the frame j.

Let us analyze the obtained distance measures, (10), (11). From (11) it is to say that the pulses contribution depends from their positions. In effect the pulses near to the right edge of the frame, give a poor contribution to the error D2, and therefore are often discarded. This fact gives rise to a residual error component which varies within the frame (buzz noise) and then, does not represent an optimal choice for the distance criteria.

From (10), we observe:

i) The first term depends only from the multipulse sequence in the frame j;

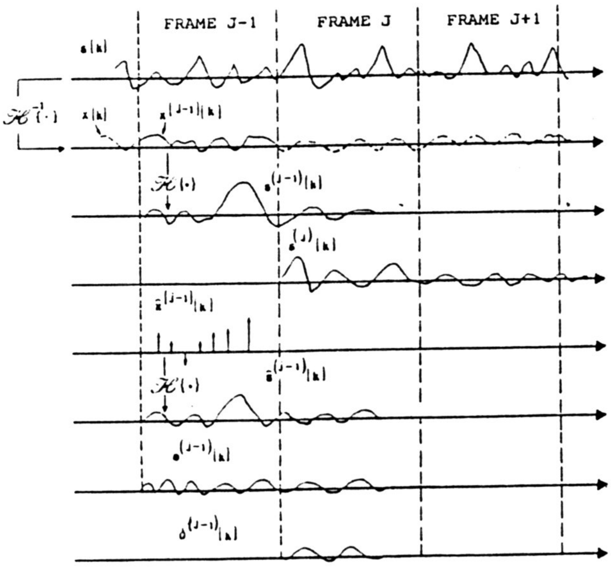

FIGURE 3 - PLOTTING IN TIME OF THE VARIOUS SIGNALS INTRODUCED IN THE TEXT

ii) The second term depends only from the multipulse sequence in the previous frame j-1;

iii) The third term depends both from $\hat{x}^{[j-1]}[k]$ and $\hat{x}^{[j]}[k]$, but is generally very small compared to the previous ones;

iv) The forth term depends only from $\hat{x}^{[j]}[k]$ and it subtracts to the energy of the signal $e_j^{(j)}[k]$ the component outside the frame interval j.

Therefore, the pulses allocated near to the right edge of the frame, give their contribution to the term $D1^{(j+1)}$, through the component ii, (that do not contribute to the choice of $\hat{x}^{[j]}[k]$) and not to the term $D1^{(j)}$, due to the subtractive component iv; in turn, this fact produces again an increase of error at the edges of each frame.

Now, observing that in the distance averaged over all frames, the terms ii) and iv) tend to be eliminated, since the components $\delta_j^{(j)}[\kappa]^2$ (from $D1^{(j)}$ ·) and $\delta_{j-1}^{(j)}[k]^2$ (from $D_j^{(j+1)}$) arise from the same term $\delta^{(j)}[k]$, filtered by weighting functions belonging to contiguous frames. This circumstance is very favourable, because it allows to erase both terms ii) and iv) in eq (11); this fact, in turn, produces a little variation in the definition of the average distance, but a noticeable improvement on the sensivity of $D1^{(j)}$ on $\hat{x}^{[j]}[k]$, and on the insensivity of $D^{(n)}$, n ≠ j, on $\hat{x}^{[j]}[k]$. As a result, our proposal is to choice $\hat{x}^{[j]}[k]$ so that the distance

$$D^{(j)} = \sum_{k=0}^{\infty} \left[ e_j^{(j)}[k] \right] + 2\sum_{k=0}^{\infty} \delta^{(j-1)}[k] \cdot e_j^{(j)}[k] \qquad 14)$$

or equivalently:

$$d^{(j)} = \sum_{k=0}^{\infty} \left\{ e_j^{(j)}[k] + \delta_j^{(j-1)}[k] \right\}^2 \tag{15}$$

is minimized. This procedure is more advantageous over those previously mentioned for the structural simplicity and the considerable elaboration saving, since it does not require contour treatments.

## 3. EVALUATION OF SIMULATION RESULTS

On the basis of the new distance criterium proposed in paragraph 2., simulation tests have been performed. The results have been compared with those obtained using previous proposed distances. In particular the distance D1 has been taken as significant benchmark. As general comment on the results is to say that with our proposed distance a better signal to coding noise ratio and a reduction of computation time have been obtained. With reference to the computation time, an important result comes relevant; infact, we have noted that neglegting the component of error $\delta^{(j-1)}[k]$ coming from previous analysis frame, only a little degradation occours, in the synthetized speech signal. Using the distance measure D1, instead, these degradations are not neglegeables. Thus, using our distance criterium a more straightforward procedure can be defined that uses only linear stationary filters. This fact leads to a relevant computational saving (more than a order of magnitude).

The curves in Figure 4 illustrate the behaviour of the segmental signal to noise ratio vs the number of pulses per frame. The first case refers to calculation without the use of the term$\delta^{(j-1)}[\kappa]$,and the other with the term $\delta^{(j-1)}[\kappa]$. The frame length has been selected to 128 samples, the preemphasis coefficient is set to 0.4, the number of LPC coefficient is 12 and $\gamma$ = 0.8. The results are obtained averanging over two sentences of male and female speaker each of 1.3 second length. Both sentences are in English language. Figure 5 illustrates the distribution of the average error along the analysis frame for our proposed distance and the distance D1. This figure confirms the results theoretically obtained, i.e. the error D1 has higher error components near the frame edge than those presented using the proposed distance criterium.

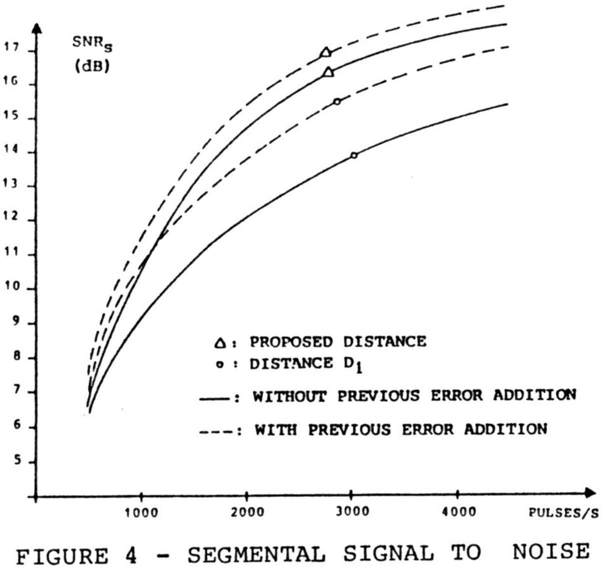

FIGURE 4 - SEGMENTAL SIGNAL TO NOISE RATIO VS PULSES PER SECOND

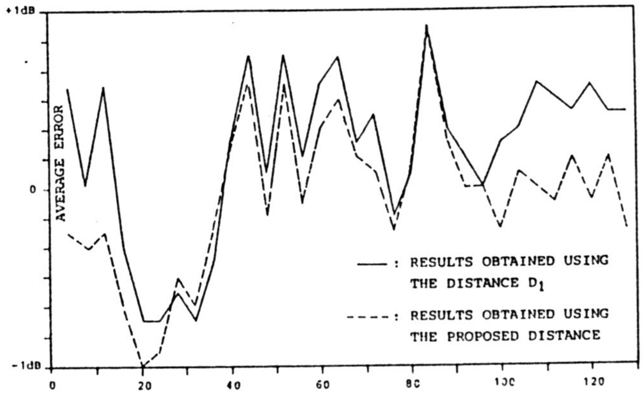

FIGURE 5 - AVERAGE ERROR VS SAMPLES POSITION IN THE ANALYSIS FRAME

## 4. CONCLUSIONS

In this paper a new procedure for calculating positions and amplitudes of pulses for multipulse voice synthetizer has been proposed. This procedure is based on a new distance criterium between original and synthetic signals, that overcomes problems due to non-stationary nature of the speech signal. Using this procedure, improvement in signal to coding noise ratio and considerable saving of calculation has been obtained in comparison with previously presented algorithms.

## REFERENCES

[1] J.D. Markel, A.H. Gray, "Linear Prediction of Speech Signal", Springer - Verlag, 1976

[2] J. Makhoul, "Linear Prediction, a Tutorial Review" Proc. IEEE Vol. 63, No 4, April 1975, pp 561 - 580

[3] B.S. Atal, J.R. Remde, "A New Model for Producing Natural Sounding Speech at Low Bit Rate", Proc. ICASSP 1982, pp 614 - 617

[4] T. Araseki, et alii, "Multi-pulse Excited Speech Coder Based on Maximum Cross-Correlation Search Algorithm", Proc. Globecom 1983, pp 23.2. 1 -23.2.5

[5] B.S. Atal, M.R. Schoeder, "Predictive Coding of Speech and Subjective Error Criteria" IEEE Trans. on Acoustic, Speech and Signal Processing, pp 274 - 249, No 6, June 1979.