# CEPSTRAL RESIDUAL VOCODER FOR IMPROVED QUALITY SPEECH TRANSMISSION AT 4.8 Kbps

D. Malah*

Electrical Engineering Dept.
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

## ABSTRACT

The cepstral representation of the LPC residual signal was found in an earlier study to provide a convenient framework for efficiently representing the main spectral components of the residual signal. The analysis-synthesis system developed uses in general poles, zeros and cepstral residual terms to better represent the input speech signal spectrum and achieves improved synthesized speech quality.

In this work we consider the quantization of the system parameters and its effect on the system performance in terms of spectral mismatch. The results are used to develop a cepstral residual vocoder system for 4.8 kbps transmission of speech. The system is based on an 8-pole LPC model and the cepstral representation of the residual signal. It applies optimal uniform quantization to each term of the cepstral residual and obtains most of the attainable spectral matching at this useful rate.

The study is based on computer simulations with telephone bandwidth speech.

## INTRODUCTION

One of the main reasons for the loss in quality of LPC synthesized speech is the spectral mismatch between the LPC model and the input speech signal. Since the residual signal contains the spectral information not extracted by the LPC analyzer, much effort was directed towards the efficient representation of the residual signal. It was noted in [1] that none of the different reported techniques (including more recent ones [2,3]) utilize all the spectral characteristics of the residual signal. such as spectral nulls. input spectral shaping effects and phase redundancy [4]. This led to the development of an analysis-synthesis system [1] which is capable of efficiently representing the spectral envelope of the residual signal. The system uses a cepstral representation of the residual signal (termed "cepstral residual") which provides a convenient framework for separating and efficiently representing the different spectral components of the residual signal. Three main spectral components are distinguished: antiresonances due to nasalized sounds; fixed or slowly varying spectral contributions which are not adequately represented by the LPC model (e.g., input spectral shaping); and the remaining spectral mismatch. An efficient representation of the first component is given by zeros which are extracted from the cepstral residual (CR) after subtraction of its long term average. Long-term averaging of each of the (CR) terms results in a bias sequence which represents the second spectral component mentioned above. The remaining spectral mismatch is represented by the remaining CR terms. In the study performed in [1] we examined the effect of using different numbers of parameters (poles, zeros and CR terms) on the reduction of the spectral mismatch - without quantization. In this work we report on the effect of quantizing the various system parameters on the system performance and develop a vocoder system which attains most of the spectral mismatch reduction at the useful rate of 4.8 kbps.

## GENERAL ANALYSIS-SYNTHESIS SYSTEM

The general block diagram of the analysis system used in this study is similar to the block diagram given in Fig. 3 of [1]. It was modified only to include the quantization of the system parameters, as follows.

The partial correlation coefficients $k_i$, $i=1,\ldots,P$, computed by the LPC analyzer are transformed to log-area ratio coefficients [5], $g_i = \log[(1-k_i)/(1+k_i)]$, and are quantized for transmission. (quantization details are given in the next section). The quantized parameters $g'_i$ are converted to the linear prediction coefficients $a'_i$ using recursive relations between $k'_i$ and $a'_i$ [5]. The complex cepstrum $\hat{h}'(n)$ of the all pole model (with quantized coefficients) is computed recursively from $a'_i$ [5]. The (real) cepstrum $c_{a'}(n)$ is then computed from $\hat{h}'(n)$ [1]. The cepstral representation, $c_r(n)$, of the residual signal is given from $c_r(n)=c(n)-c_{a'}(n)$, where $c(n)$ is the cepstrum of the input (spectrally shaped and windowed) signal.

The cepstrum of the input signal is also used to find the pitch, making voicing decisions, and compute the gain according to [1] $\text{Log } G=c(0)+P_G c(N_p)$ where $N_p$ is the pitch period and $P_G$ is a constant [1] (we used $P_G = 2.2$).

To remove the effect of pitch, only $N_c<N$ cepstral residual (CR) terms are used for spectral matching. This corresponds to multiplying $c_r(n)$ by a rectangular window $w_c(n)$ [1]. The windowed CR is used to compute the long-term average $\langle c_r(n)\rangle$ (a bias sequence) which is subtracted from $c_r(n)$ resulting in the unbiased CR - $\tilde{c}_r(n)$. The bias sequence $\langle c_r(n)\rangle$ is the cepstral representation of the fixed input spectral shaping which was not

* This work was carried out at the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, USA, while on leave from the Technion.



CH 1746-7/82/0000 - 0622 $ 00.75 © 1982 IEEE

removed by the all-pole model. The subtraction of $\langle c_r(n)\rangle$ from $c_r(n)$ corresponds to inverse filtering in the frequency domain which removes the effects of the input spectral shaping [1]. For systems with non-fixed input conditions, such as in the switched telephone network, the tracking of the bias sequence is useful and can be done recursively using $\langle c_r(n)\rangle_m = \alpha\langle c_r(n)\rangle_{m-1} + (1-\alpha)(c_r(n))_m$, where m denotes the analysis frame number and $\alpha$ is a constant, $0<\alpha<1$, which determines the update rate (a time constant of 1 second appears reasonable). The updated bias sequence can be transmitted during short silence intervals or during unvoiced segments using the pitch data slot.

In the system proposed later for 4.8 kbps transmission by speech, the unbiased CR - $\tilde{c}_r(n)$ is quantized and transmitted. However, in view of the results in [1] (without quantization) we examined also the possibility of computing zeros from the unbiased CR and transmitting them in quantized form as explained below. As elaborated in [1], once the fixed spectral shaping effects were removed, by subtracting $\langle c_r(n)\rangle$ from $c_r(n)$, the representation of the remaining spectral mismatch by zeros is much more efficient than reported in earlier works (e.g., [6]) since only few zeros (typically 4) are needed now to represent antiresonances in the given speech band (telephone bandwidth) and the remaining spectral mismatch is represented by the remaining CR terms $\tilde{c}_r(n)$ [1]. The zeros are found by applying an LPC analysis to the impulse response representing the inverse spectral envelope of the residual signal [1]. Following the computation of the partial correlation coefficients $k_z(i)$, representing the zeros, these coefficients are transformed to log-area ratio coefficients $g_z(i)$, which are quantized and transmitted. The remaining CR, $\tilde{\tilde{c}}_r(n)$, are computed by subtracting from $\tilde{c}_r(n)$ the cepstral representation of the zeros - $c_{b'}(n)$. The remaining CR terms $\tilde{\tilde{c}}_r(n)$ are weighted by a Hanning window, $w_r(n)$, (to provide additional smoothing [1]), are quantized and transmitted.

The synthesizer block diagram is the same as in Fig. 4 of [1]. It is based on the usual LPC synthesizer with the addition of zeros and spectral shaping of the excitation.

For illustration, Fig. 1 shows the improvement obtained in spectral matching by using 4 zeros and 11 cepstral residual peaks [1] to represent the residual signal of an 8 pole LPC system (unquantized). The remaining spectral mismatch in the example in Fig. 1 is $E_c$=0.9 dB which is an RMS Log-spectral measure defined in [1].

## PARAMETER SELECTION AND QUANTIZATION

The effect of selecting various numbers of poles, zeros, and cepstral residual (CR) terms (without quantization) on the spectral mismatch reduction was examined in detail in [1]. One important result is that because additional parameters are used to represent the residual signal, it is sufficient to use 8 poles in the LPC model (for telephone bandwidth speech). Another result is that CR peaks are more efficient for representing the residual then sequentially ordered cepstral terms [1]. However, when we considered the quantization of these parameters, we found that the use of sequentially ordered CR terms is preferable, because encoding the location of the cepstral peaks requires too many bits. Yet another result is that the number of sequentially ordered CR terms needed to obtain a given spectral mismatch reduction is smaller by up to 20% than the number of zeros needed for the same mismatch reduction. One can conclude therefore that only CR terms should be used. However, the zeros were found to provide a better representation of sharp spectral nulls and in informal listening we perceived an additional small improvement for nasal sounds when zeros were also included. We have considered therefore the following two parameter selections:

(1) 8 poles and sequentially ordered CR terms (to be denoted as selection PCR).

(2) 8 poles, 4 zeros, and sequentially ordered CR terms (selection PZCR).

Examination of the effect of quantizing the poles and zeros on the spectral mismatch, when CR terms are also used, revealed that if 32 bits are used for representing the poles and 10 bits for representing the zeros, the effect of this quantization is very small. The bit assignments used for the log-area ratio parameters were (5,5,5,4,4,4,3, 2) and (3,3,2,2), respectively. These bit assignments are based on minimizing the mean square quantization error (as explained below for the CR terms). Six sentences were used (3 by male and 3 by female) in computing the average spectral matching.

To study the effect of quantizing the CR terms we have first considered the approach of encoding all the chosen terms with an equal number of bits. Uniform quantizers were used with optimal step sizes matched to its amplitude histograms which were found to be different not only in variance but also in shape. The optimal step size was computed for each term to minimize the mean square quantization error. The computation was done on the unwindowed terms and the windowing was applied later to the step sizes as well. This was needed since for female voice the window length may vary with the pitch period. The solid lines in Fig. 2 show the results obtained for the above parameter selections, PCR and PZCR, respectively. The figure shows the average spectral mismatch obtained for a given total number of bits assigned to the CR terms used for spectral matching. Each point on these curves correspond to a specific bit assignment which gives the minimum average spectral mismatch. The number of bits/term assigned in some specific cases is shown. Note that the total number of bits for transmission is different for the PCR and PZCR selections since additional 10 bits (per frame) are needed for representing the 4 zeros.

Next, we consider a bit assignment which for a given total number of bits, $B_c$, allocated to the cepstral residual terms, minimizes the total mean square quantization error. This also minimizes the effect of quantizing the CR terms on the system performance in terms of $E_c$. The solution to this optimal bit assignment problem is well known and was applied in dynamic bit allocations (from frame to frame) for adaptive transform coders [7,8]. The solution is that the mean square quantization error of each term must be the same. Since here the bit

assignment is fixed and not dynamic, we used the graphical approach elaborated in [8] to obtain different bit assignments. According to this approach, we have plotted the variance (in dB) of the CR terms, as shown in Fig. 3, and super-imposed on it a set of parallel threshold lines $\lambda_i$, separated apart by $\lambda$=5dB. (See also Fig.13 in [8]). Terms with variance values between $\lambda_i$ and $\lambda_{i-1}$ receive i bits. It is noted that terms which have a variance below $\lambda_0$ are not transmitted (0 bits) and terms with variances above a given threshold line, say $\lambda_k$, are all given k bits. By considering different integer values for k, and moving the set of threshold lines up or down, one obtains different bit assignments. The value of $\lambda$=5dB is used because we found in simulations that the addition of one bit to each of the optimal uniform quantizers resulted in a reduction of about 5dB in the mean square quantization error (in the range of 1 to 5 bits). Using this method, different optimal bit assignments were found and were used to quantize the CR terms. The remaining spectral mismatch was computed from the difference between the original CR sequence and the quantized sequence. The dashed lines in Fig. 2 show the simulation results for the parameter selections PCR and PZCR, respectively.

Taking into account that 10 bits are needed to represent the 4 zeros used in the PZCR selection, we find from Fig. 2 that the PCR selection gives a lower spectral mismatch than the PZCR selection for the same total number of bits used to represent the residual signal. Combining this result with the fact that the marginal improvement obtained by using also zeros happens only for a relatively small number of speech segments (having sharp spectral nulls), and adding to it the reduction in implementation complexity, we consider the PCR selection as more attractive and useful. More details on the vocoder system using this selection are given in the next section.

Before we end this section we would like to mention that we considered also using differential coding (DPCM) of the CR terms to further reduce the bit rate. For that purpose we measured the average correlation between cepstral terms in the same frame and also the correlation for each term from frame to frame. The intraframe normalized correlation was found to be very small ($\rho(1) = -0.025$). The interframe correlation is higher with average values of $\rho_i(1)$= 0.6,0.4,0.2 for i=1,15,30, respectively, where i is the index of the CR term. We did not consider these values to be sufficiently high to justify differential coding in this system.

## CEPSTRAL RESIDUAL VOCODER

From the results of the previous section the following vocoder system emerges. The transmitter consists of the analysis system shown in Fig. 4. Its parameters are the coefficients representing the poles and the cepstral residual terms. No zeros are used. The quantization of the windowed $\tilde{c}_r(n)$ is done according to the optimal bit assignment described in the previous section using an overall number of $B_c$ bits. The LPC analyzer extracts 8 poles and the log-area ratio coefficients representing the poles are assigned 32 bits. Allowing 13 bits for pitch, gain, V/UV and sync and having 50 frames/sec, we obtain that the transmission bit rate is given by $(2250+50B_c)$ bps. Using the dashed curve (PCR) in Fig. 2 (which gives the spectral mismatch as a function of $B_c$) we obtain the curve in Fig. 5 which shows the spectral mismatch as a function of the transmission bit rate. Of particular interest is the rate of 4800 bps, since at this useful rate the spectral mismatch is sufficiently reduced to provide sufficiently improved speech quality (over LPC alone) as to justify the use of this more complex system. For this rate $B_c$=51 bits and the bit assignment used is $(3^8,2^9,1^9)$, where $b^m$ means m repetitions of b.

The receiver consists of the synthesizer shown in Fig. 6. In this figure, $c_t(n)$ is the received CR sequence. The figure does not show the interpolation of the parameters and the conversion of the g-parameters to LPC.

## CONCLUSION

One of the motivations for using the cepstral representation of the residual signal was the ability of conveniently extracting the zeros. The cepstral residual vocoder that emerged from this work does not use zeros since the cepstral residual terms alone were found to more efficiently represent the overall spectral mismatch. Still, this representation is justified in view of the advantages it offers over directly encoding the spectral envelope in the log-frequency domain. These include the simple removal of the effect of pitch (by cepstral windowing); the convenient and efficient representation of input spectral shaping (bias); The efficient representation of the unbiased spectral envelope with a relatively small number of uncorrelated cepstral residual terms; an inherent gain normalization which avoids the need for adaptive quantizers (c(0) contains the gain information); and finally, the cepstral representation also provides robust pitch detection.

The proposed approach has also an advantage on using homomorphic prediction [6], in which the LPC coefficients (poles) are found from the cepstral representation. The reason is that the cepstral window used to remove the pitch causes undesired smearing of high Q formant peaks.

The presented 4.8 kbps cepstral residual vocoder was not fully tested yet. However, informal listening has shown an improvement in quality over LPC alone which well justifies its use - if a computation capability exists and if doubling the LPC 2.4 kbps transmission rate is acceptable.

An important remaining source of degradation is the need for V/UV decisions. While systems that encode the residual signal in the time domain (and do not explicitly extract pitch and voicing information) overcome this problem, they are less efficient in representing the spectral envelope and are unable to utilize the phase redundancy in the residual signal.

## REFERENCES

[1] D. Malah, "Efficient spectral matching of the LPC residual signal", Proc. IEEE Int. Conf. ASSP, pp. 1288-1291, 1981.

[2] B.M. Abzug, "Using the prediction residual to improve LPC synthesis for 9600 bps applications", Proc. IEEE Int. Conf. ASSP, pp. 812-815, 1981.

[3] H. Katterfeldt, "A DFT-based residual-excited linear predictive coder (RELP) for 4.8 and 9.6 kb/s", Proc. IEEE Int. Conf. ASSP, pp. 824-827, 1981.

[4] B.S. Atal and N. David, "On synthesizing natural-sounding speech by linear prediction", Proc. IEEE Int. Conf. ASSP, pp. 44-47, 1979.

[5] L.R. Rabiner and R.W. Schafer, "Digital processing of speech signals", Prentice Hall, New Jersey, 1978.

[6] G.C. Kopec et al., Speech analysis by homomorphic prediction", IEEE Trans. ASSP, Vol. ASSP-25, No. 1, pp. 40-43, Feb. 1977.

[7] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals", IEEE Trans. ASSP, Vol. ASSP-25, pp. 299-309, Aug. 1977.

[8] J.M. Tribolet and R.E. Crochiere, "Frequency domain coding of speech", IEEE Trans. ASSP, Vol. ASSP-27, pp. 512-530, Oct. 1979.

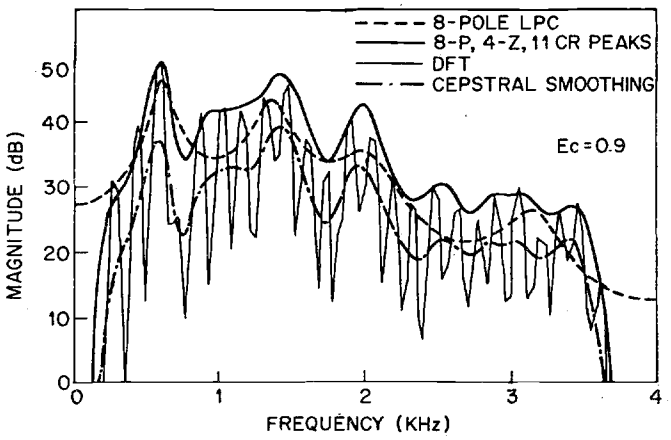


Fig. 1 Spectral matching of the spectral envelope.

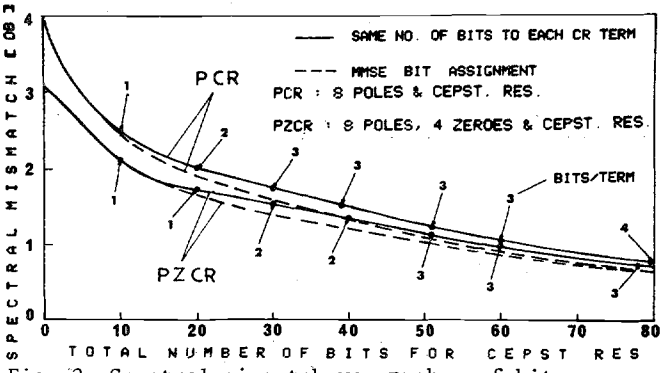


Fig. 2 Spectral mismatch vs. number of bits assigned to the cepstral residual terms.

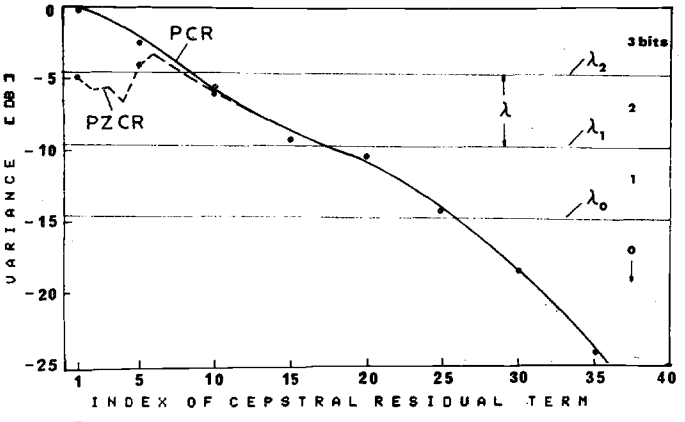


Fig. 3 Variance of cepstral residual terms.

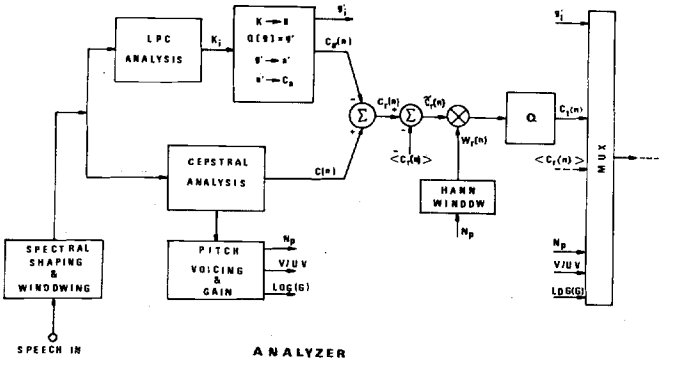


Fig. 4 Block diagram of analyzer.

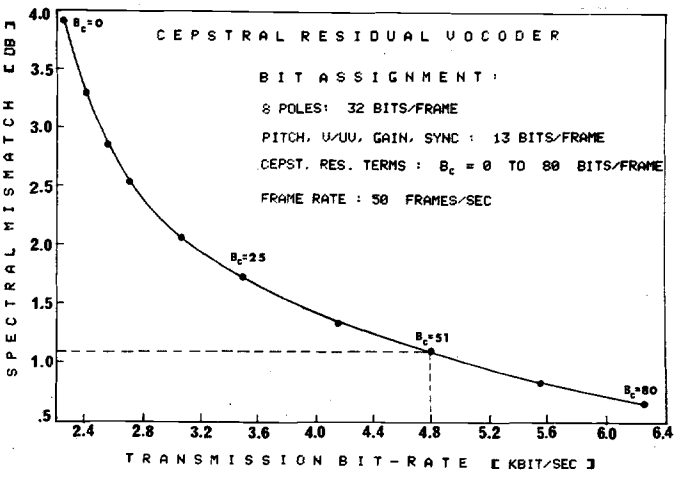


Fig. 5 Spectral mismatch vs. transmission bit-rate.

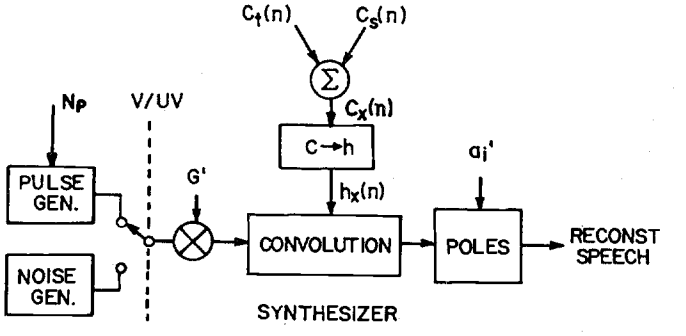


Fig. 6 Block diagram of synthesizer.