EFFICIENT SPECTRAL MATCHING OF THE LPC RESIDUAL SIGNAL

*D. Malah**

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

A basic reason for the loss of quality in LPC synthesized speech is the spectral mismatch between the LPC model and the analyzed speech. Since the residual signal contains the spectral information not extracted by the LPC analyzer, existing methods for improving the synthesized speech quality are based on encoding the residual signal or reducing its information content by improved pole-zero models. In this work we study an approach in which the spectral envelope of the residual signal is divided into three main components which are separately represented to provide an overall more efficient representation of the residual signal. The three spectral components are the antiresonances due to nasalized sounds, fixed or slowly varying spectral contributions due to input spectral shaping, and the remaining spectral matching error. An analysis-synthesis system based on the cepstral representation of the residual signal is developed and is shown to provide a particularly convenient framework for separating and representing by zeroes and cepstral residuals the different spectral components. Preliminary results of the study are presented on the basis of simulations with telephone bandwidth speech.

INTRODUCTION

Linear predictive coding (LPC) is an efficient and widely used technique for low bit rate transmission of speech. However, the quality of the LPC synthesized speech is speaker and environment dependent and is not natural sounding. A basic reason for the loss of quality is the spectral mismatch between the spectrum of the all-pole LPC model and the input speech spectrum [1]. Since the residual error signal contains the spectral information not extracted by the LPC analyzer, efforts for improving the synthesized speech quality were centered on the efficient representation and encoding of this signal, e.g. [2-6], as well as the reduction of its information content by improved pole-zero models, e.g. [7-9]. In a related effort, the modification of the buzz-hiss excitation source was also considered [10,11]. While the antiresonances in nasalized sounds can be adequately represented with a relatively small number of zeroes, other sources of spectral mismatch such as the glottal pulse, input spectral shaping, and other deviations from the assumed rational model, typically require a large number of zeroes for adequate spectral matching. The modification of the LPC buzz-hiss excitation model [10,11] was reported to reduce the characteristic buzziness in LPC synthesized speech but can not sufficiently reduce the spectral mismatch as to render the LPC system robust and natural sounding. As for residual encoding, since standard waveform coding of the full band residual signal cannot be afforded in low bit rate applications, one approach is to encode the residual signal baseband [2,3]. At the receiver, the residual baseband signal is used to generate a full band excitation signal by using high frequency regeneration techniques [12]. In addition to the limited spectral matching the regeneration process causes perceivable distortions. Recent works have avoided the above problem by encoding the full band residual signal using time and frequency domain techniques. In the time domain, center clipping and fine quantization of high amplitude residual samples were applied [6,13]. In addition, replication of a pitch period residual was also found useful [6]. In the frequency domain, subband coding of the residual signal [4] and the coding of its spectral envelope [5], were considered. The latter was based on an experiment which has shown that, perceptually, the spectral magnitude information in the residual signal is significantly more important than the phase

From the above review of existing techniques we note that each technique exploits some properties and redundancies in the residual signal, but is unable to take advantage of all of them. For example, time-domain techniques cannot exploit directly the phase redundancy reported in [5], and are not as efficient in representing antiresonances as zeroes. The use of zeroes only is not as efficient for the representation of other spectral components in the residual, and encoding of the spectral envelope is not as efficient as zeroes for representing antiresonances. In this work we have taken therefore the approach of dividing the spectral envelope of the residual signal into three main components, so that each component can be represented separately in a more efficient way. The three spectral components are the antiresonances due to nasalized sounds; fixed or slowly varying spectral contributions which are not adequately represented by the all-pole LPC model and can be due to input spectral shaping (e.g. filtering), speaker characteristics, and effects of environment conditions; and the remaining spectral matching error. An analysis-synthesis system based on the cepstral representation of the residual signal is developed and is shown to provide a particularly convenient framework for separating and representing by zeroes or cepstral residuals each of the above spectral components.

CEPSTRAL ANALYSIS OF THE RESIDUAL SIGNAL

The cepstral analysis [14] of the residual signal was considered in this work since it provides means for smoothing the spectral envelope, extraction of zeroes [8,15], removal (deconvolution) of input spectral shaping effects, and robust pitch detection [16].

The usual way for obtaining the LPC residual signal is to inverse filter the input signal using the extracted LPC parameters. The cepstral representation of the residual signal, to be called the *cepstral residual*, is then computed from this signal. However, to save computations and reduce the delay involved in first extracting the LPC parameters and then inverse filtering, the scheme shown in Fig. 1 was used. According to this scheme, the LPC and cepstral analyses are both performed on the pre-emphasized and windowed input speech signal. The LPC parameters are then used to compute N_c cepstral components $c_a(n)$ ($N_c < N_p$ – the pitch period duration in samples) of the all-pole model, through well known recursive relations [17].

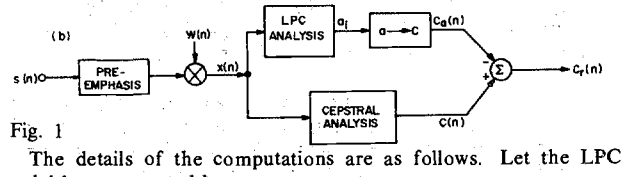

Fig. 1

The details of the computations are as follows. Let the LPC model be represented by

$$H(z) = G_A/A(z) = G_A/(1 - \sum_{i=1}^{p} a_i z^{-i}) \qquad (1)$$

The complex cepstrum is defined as the inverse Fourier transform of the complex logarithm of $H(e^{j\omega})$ (the frequency response of the all-pole model). Since $H(z)$ is minimum phase its

*On leave from the Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa, Israel.

CH1610-5/81/0000-1288 $00.75

complex cepstrum $\hat{h}(n)$ is one sided, i.e. $\hat{h}(n) = 0$ for $n < 0$, and can be computed from the LPC parameters by [17],

$$\hat{h}(n) = \begin{cases} a_n + \sum_{k=1}^{n-1} (k/n)\hat{h}(k)a_{n-k}, & 2 \le n \le p \\ \sum_{k=1}^{p} ((n-k)/n)\hat{h}(n-k)a_k, & p < n \end{cases} \quad (2)$$

with $\hat{h}(0) = Log(G_A)$ and $\hat{h}(1) = a_1$. Since we are interested in the spectral magnitude information, we will use in the sequel the (real) cepstrum instead of the complex cepstrum. The cepstrum is an even function defined as the inverse Fourier transform of $Log\,|H(e^{j\omega})|$ which is the real part of the complex logarithm of $H(e^{j\omega})$. Hence, given $\hat{h}(n)$, the cepstrum $c_a(n)$ is given by $[\hat{h}(n) + \hat{h}(-n)]/2$. Finally, since $\hat{h}(n)$ is one sided, we have

$$c_a(0) = \hat{h}(0) \text{ and } \quad c_a(n) = \hat{h}(|n|)/2\,, \; n \neq 0 \quad (3)$$

The cepstrum of the pre-emphasized and windowed speech signal $x(n)$ is denoted by $c(n)$ and is computed by inverse Fourier transforming $Log\,|X(e^{j\omega})|$. The cepstral residual $c_r(n)$ is thus given by $c_r(n) = c(n) - c_a(n)$.

We have found in our study that for an efficient representation of the antiresonances by zeroes it is extremely important to remove first the effects of input spectral shaping, such as bandpass filtering, which may require many zeroes for its representation. We address therefore this issue first.

Removal of Input Spectral Shaping Effects: Since the cepstral representation of a signal is not a parametric one and describes the signal as is, $c_r(n)$ contains the cepstral representation of the fixed input spectral shaping (which was not removed by the all-pole model) as a bias. This bias can be estimated by computing a long-term average $<c_r(n)>$ of $c_r(n)$. Subtraction of $<c_r(n)>$ from $c_r(n)$ corresponds to inverse filtering in the frequency domain which removes the effects of the input spectral shaping and provides for a more efficient representation of the unbiased residual $\tilde{c}_r(n)$, where $\tilde{c}_r(n) = c_r(n) - <c_r(n)>$. In systems with varying input conditions, slow tracking of the bias sequence $<c_r(n)>$ is useful and can be done by using a running average. The updated bias sequence can be transmitted at a slow rate during silence periods or even during unvoiced segments by using the pitch slot.

Computation of Zeroes: The scheme for computing the zeroes from $\tilde{c}_r(n)$ is shown in Fig. 2. The rectangular window $w_c(n)$ corresponds to using only the first N_c cepstral terms. As in homomorphic prediction [8,15] the cepstrum is used to find a minimum phase impulse response to which the LPC analysis is applied. In this case, the spectrum is first inverted by inverting the sign of $\tilde{c}_r(n)$. To compute the minimum phase impulse response $h_z^{-1}(n)$ of the inverted spectrum, the one sided complex cepstrum $h_z^{-1}(n)$ is first found from $\tilde{c}_r(n)$, using the relations discussed in conjunction with eq. (3). The minimum phase impulse response $h_z^{-1}(n)$ is modeled by $H_z^{-1}(z)$, which takes the form of (1) but with b_i replacing a_i (and proper gain). The coefficients b_i are used to represent the zeroes.

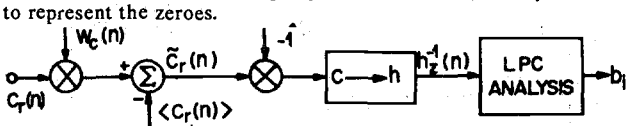

Fig. 2 Computation of zeroes from cepstral residual

The Remaining Cepstral Residual: Following the modeling by zeroes, the remaining cepstral residual $\tilde{\tilde{c}}_r(n)$ can be computed in a similar fashion to the computation of $c_r(n)$. This is done by using again the relations in (2), with b_i replacing a_i and sign inversion, to find the cepstral representation of the zeroes, $c_b(n)$, and then $\tilde{\tilde{c}}_r(n) = \tilde{c}_r(n) - c_b(n)$.

If desired, components of $\tilde{\tilde{c}}_r(n)$ can now be selected for transmission. However, we found it useful to first window $\tilde{\tilde{c}}_r(n)$ by a suitable cepstral window $w_r(n)$ which further smoothes the log-spectrum of the remaining residual. In the sequel, the windowed remaining cepstral residual is denoted by $c_{rw}(n)$ and its components which are selected for transmission by $c_t(n)$.

SPECTRAL MISMATCH MEASURES

Let $Log|E(e^{j\omega})|$ be the Fourier transform of a given cepstral residual sequence $c_e(n)$. Then, by Parseval's relation,

$$\epsilon_c^2 \triangleq \sum_{n=-N_c}^{N_c} c_e^2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} [Log|E(e^{j\omega})|]^2 \, d\omega \quad (4)$$

A useful evaluation of the spectral mismatch is given in terms of the RMS Log-spectral mismatch in dB:

$$E_c \triangleq [\epsilon_c^2]^{1/2} \; 20/Log(10) \; [dB] \quad (5)$$

It is of interest to point out that E_c is minimized for a given number N_r of cepstral residual coefficients if the N_r cepstral components having the largest magnitude are selected.

Another approach for measuring the spectral mismatch is related to the prediction error measure used in LPC analysis, [7] The normalized prediction error V_p is defined as the ratio between the energy of the LPC residual signal to the energy of the input signal E_x. It was established that [7] $V_p = G_A^2/E_x$, where G_A is the gain term in (1). It was also shown that the minimum value of V_p is given by [7] $V_{\min} = G_c^2/E_x$, where G_c is given from the input signal cepstrum $c(n)$ by $G_c = \exp(c(0))$. It can be shown that

$$V_p = V_{\min} \sum_n \bar{h}_r^2(n) \quad (6)$$

where $\bar{h}_r(n)$ is a normalized impulse response sequence obtained from $c_r(n)$ by setting first $c_r(0)$ to zero.

Clearly, if there is no residual error, $\bar{h}_r(n)$ becomes a unit impulse and $V_p = V_{\min}$. In our application $\bar{h}_r(n)$ is the impulse response needed to shape the synthesizer excitation signal in order to compensate for the matching error introduced by the all-pole LPC model. In view of (6) we define a spectral mismatch measure

$$E_h \triangleq 10 \; Log_{10} [\sum_n \bar{h}^2(n)] \; [dB] \quad (7)$$

where $\bar{h}(n)$ corresponds to a given cepstral residual sequence. Both measures, E_c and E_h, can be computed at any given stage of the matching process by using the given cepstral residual.

ANALYSIS-SYNTHESIS SYSTEM

The general block diagram of the analyzer is shown in Fig. 3. It is based on the spectral matching process described above and contains the blocks described in Fig.'s 1 and 2. Additional blocks shown in Fig. 3 are for pitch detection and voicing decision [16], gain computation, cepstral residual selection and the quantization and encoding of transmitted parameters.

The optimal gain is $G_c = \exp(c(0))$. However, since the matching is done only for N_c cepstral coefficients, G_c is too small. The most pronounced component beyond N_c is the cepstral peak at N_p - the pitch period. The proper gain G is obtained by modifying G_c with the gain due to the pitch peak. In the Log domain we have $Log\ G = c(0) + 2c(N_p)$.

The synthesizer can be realized in several ways. The block diagram of one possible scheme is shown in Fig. 4. This scheme uses a pole-zero synthesizer which is driven by an excitation signal generated by convolving the pitch pulses or noise with the impulse

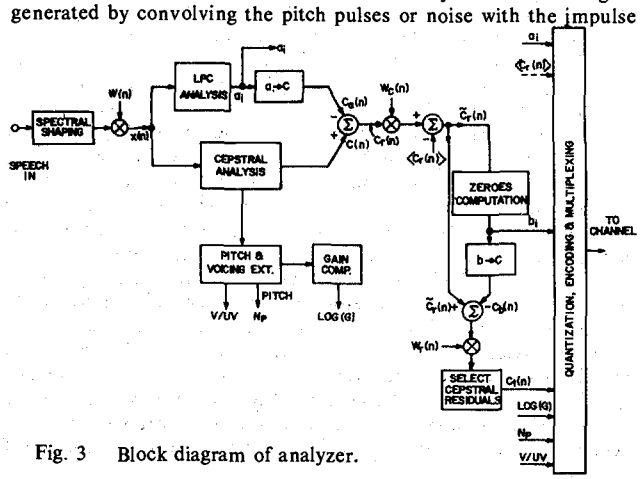

Fig. 3 Block diagram of analyzer.

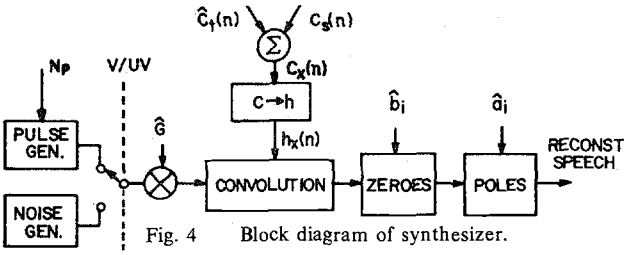

Fig. 4 Block diagram of synthesizer.

response $h_x(n)$. This impulse response is computed from $c_x(n)$ which is the sum of the received cepstral residual $\hat{c}_f(n)$ and a cepstral shaping sequence $c_s(n)$. $c_s(n)$ includes the bias sequence $<c_r(n)>$ as well as any other desired spectral shaping such as deemphasis, band-edge enhancement to remove the smoothing effect of the cepstral window, etc. One alternative to this scheme is to use an all-pole LPC synthesizer and include the zeroes, through their cepstral representation, $\hat{c}_b(n)$, in $c_x(n)$. Another alternative is to use a homomorphic synthesizer in which $c_x(n)$ includes the cepstral representations of the poles and zeroes.

PRELIMINARY SIMULATION RESULTS

The analysis-synthesis system described above was computer simulated. The results reported here are, however, preliminary and limited since they are based on several utterances only spoken by three male and one female, and do not yet include quantization of the parameters. The input speech (both by telephone and microphone) was recorded and sampled at 8 kHz and was digitally highpass filtered to well define the input lower band edge (300 Hz). This was found important for designing the spectral shaping at the synthesizer (needed for compensating the effect of cepstral smoothing on the lower band-edge). The filtered speech was preemphasized by a first order differencing operation ($1-0.95z^{-1}$) and windowed by a Hanning window ($w(n)$) of 256 samples duration (32 msec). The data update was 16 msec. Following the cepstral analysis and the computation of the cepstral representation of the LPC model ($c_a(n)$), the raw cepstral residual $c_r(n)$ was computed but $c_r(0)$ was set to zero since the gain is computed separately, as discussed earlier. The number of cepstral residual terms N_c was chosen to be 50 if $N_p \geq 52$, and $N_c = N_p - 2$ if $N_p < 52$. Thus, for male speakers N_c is usually fixed and is equal to 50, whereas for female speakers N_c depends on the pitch period N_p. Next, the long term average $<c_r(n)>$ is computed. Fig. 5 shows typical forms of $c_r(n)$, $<c_r(n)>$, and the unbiased residual $\tilde{c}_r(n)$. The spectral representation of each is shown in Fig. 6. The effect of increasing the number of poles, and of removing the bias sequence $<c_r(n)>$, on the spectral measures E_c and E_h is shown in Fig. 7. The results show that beyond 12 poles the spectral mismatch decreases very slowly with the increase in the number of poles. In addition, it is seen that the bias contributes significantly to the spectral mismatch. Since the bias represents fixed or slowly varying spectral shaping its use by itself is not expected to significantly improve the speech quality. This was verified by informal listening in which a certain reduction in buzziness and some increase in crispiness was noted but not commensurate with the large decrease in mismatch seen in Fig. 7. The remaining unbiased mismatch is removable by using zeroes and/or the remaining cepstral residual. Its removal was found to significantly contribute to the quality of the synthesized speech. It is concluded therefore that a meaningful mismatch measure should be based on the unbiased residual. The unbiased cepstral residual $\tilde{c}_r(n)$ was used to extract the zeroes. Simulations have shown that once the bias is removed a low order model is sufficient to represent the antiresonances in the speech spectrum. With two zeroes per antiresonance, 4 zeroes are sufficient to represent the maximum of two antiresonances expected in the given bandwidth. Fig.'s 5b and 6b show $c_b(n)$ and its transform, respectively, using 4 zeroes which are computed from $\tilde{c}_r(n)$ of Fig. 5b. The number of zeroes used need not to be limited to 4 since by increasing its number further reduction in spectral mismatch is achieved. The mismatch reduction was not however as effective if

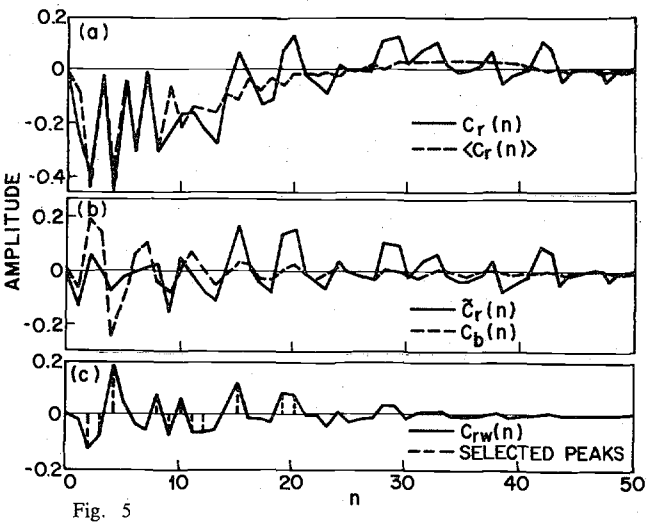

Fig. 5

the bias is not removed first. The effect of increasing the number of zeroes on the reduction of the spectral mismatch is shown in Fig. 8 for both 8 and 12 poles. The effect of the bias removal is clearly seen. Note also that once the bias is removed the use of 12 poles gives very little improvement over 8 poles, especially if the number of zeroes is sufficiently large. It is seen from Fig.'s 7 and 8 that the behavior of the two measures E_c and E_h is very similar. Since E_c is easier to compute we preferred its use in the subsequent simulations. We next examined the reduction of the remaining spectral mismatch by selecting components ($c_i(n)$) from the windowed cepstral residual $c_{rw}(n)$ ($w_r(n)$ was a Hanning window with zero value at N_c). As an example, Fig. 5c shows $c_{rw}(n)$ obtained from $\tilde{c}_r(n)$ and $c_b(n)$ of Fig. 5b, and the selection of 11 peaks. The corresponding frequency transforms are shown in Fig. 6c. The difference between the solid and dashed line in Fig. 6 is the final matching error. Fig. 9 presents the mismatch E_c obtained with an 8 pole LPC model and a varying number of zeroes for different selections of cepstral residual components. Two different cepstral residual selections were considered. An ordered (sequential) selection of cepstral coefficients (solid lines) and the selection of cepstral residual peaks (dashed lines).

Fig. 9 is valuable in selecting different alternatives for representing the LPC residual signal. For a given desired mismatch (in dB), different selections of zeroes and cepstral residuals (ordered or peaks) can be used to yield the same mismatch. The final choice depends on the number of bits needed to represent the different

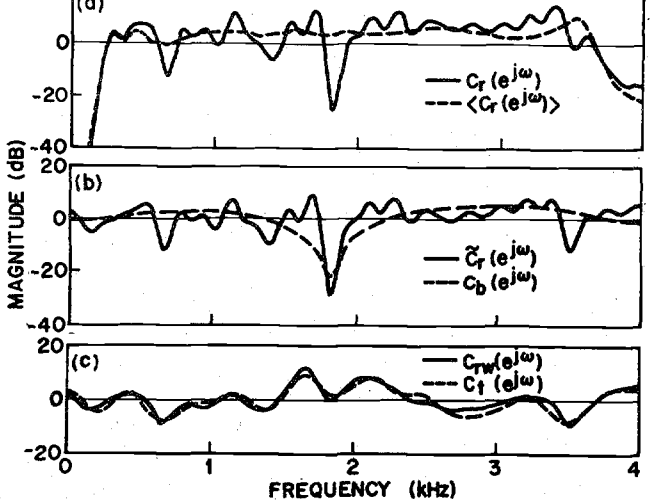

Fig. 6 Fourier transforms of the cepstral residuals

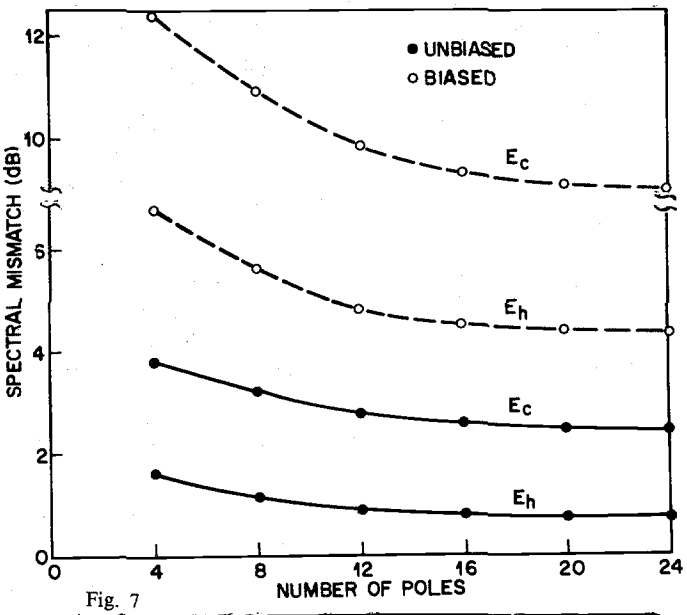

Fig. 7

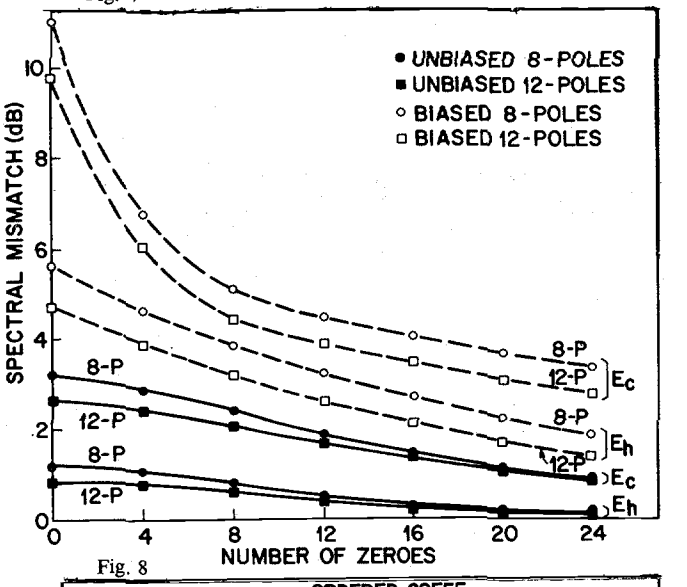

Fig. 8

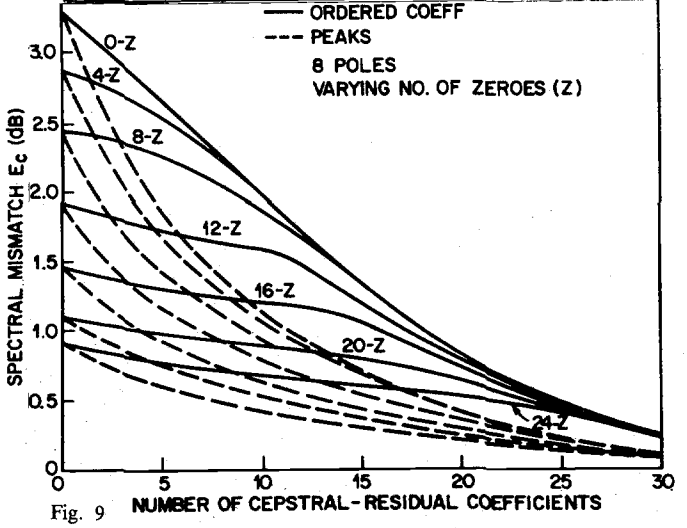

Fig. 9

coefficients (the quantization, if not fine enough, may also alter the figure). The issue of quantization is now under study. It is worthwhile to note that in the simulation conducted with unquantized parameters a choice of parameters which resulted in a mismatch E_c of below 1 dB was hardly distinguishable (in informal listening) from the case where the full cepstral residual was used ($E_c = 0$). If one sets as a goal to achieve a mismatch of 1 dB, one finds from Fig. 9 the following three possible parameter choices (with the assumption that the bias $<c_r(n)>$ was removed): (a) 8 poles, 22 zeroes (b) 8 poles, 4 zeroes (for antiresonances representation), and 18 ordered cepstral residual coefficients (c) 8 poles, 4 zeroes, and 11 cepstral residual peaks. The final choice depends on the quantization effects and the number of bits required for each alternative.

CONCLUSION

An analysis-synthesis system which exploits the phase redundancy and other spectral characteristics of the LPC residual signal was developed. The separation and representation of three main spectral residual components are conveniently and efficiently obtained within a framework based on the cepstral representation of the residual signal. Preliminary simulation results with telephone bandwidth speech indicate the particular importance of removing the long-term average of the cepstral residual (i.e. the cepstral bias due to input spectral shaping) for obtaining an efficient representation of the residual by zeroes and/or cepstral coefficients.

Since quantization effects were not studied fully as yet, no recommendation is made on the final choice of parameters. However, it appears that 8 poles and 4 zeroes can be used for adequate representation of spectral peaks (formants) and dips (antiresonances) in the given bandwidth. The remaining parameters consist of either additional zeroes and/or cepstral residual coefficients. In view of the results obtained so far it is believed that the proposed system can be suitable for low bit rate applications and offers meaningful improvements in robustness and quality. Further study on a larger data base is needed however to examine its full potential.

REFERENCES

[1] D. Y. Wong, "On understanding the quality problems of LPC Speech," Proc. IEEE Int. Conf. ASSP, pp. 725-72, 1980.

[2] C. K. Un and D. T. Magill, "The residual excited linear Prediction Vocoder with transmission rate below 9.6 K bits/s," IEEE Trans. Communication, Vol. COM-23, pp. 1466-1473, Dec. 1975.

[3] M. D. Dankenberg and D. Y. Wong, "Development of a 4.8-9.6 Kbps RELP Vocoder," Proc. IEEE Int. Conf. ASSP, pp. 554-557, 1979.

[4] L. L. Burge, and R. Yarlagadda, "An efficient coding of the prediction residual," Proc. IEEE Int. Conf. ASSP, pp. 538-541, 1979.

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

[6] S. Maitra and C. R. Davis, "Improvements in the classical model for better speech quality," Proc. IEEE Int. Conf. ASSP, p. 23-26, 1980.

[7] J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.

[8] G. C. Kopec et. al., "Speech Analysis by homomorphic prediction," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 1, pp. 40-43, Feb. 1977.

[9] K. Steiglitz, "On the simultaneous estimation of poles and zeroes in speech analysis," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, No. 3, pp. 229-234, June 1977.

[10] M. R. Sambur et. al., "On reducing the buzz in LPC synthesis," J. Acoust. Soc. Am., Vol. 63, No. 3, pp. 918-924, March 1978.

[11] J. Makhoul et. al., "A mixed-source model for speech compression and synthesis," Proc. IEEE Int. Conf. ASSP, pp. 163-166, 1978.

[12] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," Proc. IEEE Int. Conf. ASSP, pp. 428-431, 1979.

[13] B. S. Atal and M. R. Schroeder, "Improved quantizer for adaptive predictive coding of speech signals at low bit rate," Proc. IEEE Int. Conf. ASSP, pp. 535-538, 1980.

[14] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Inc. Englewood Cliffs, N.J. 1975.

[15] A. V. Oppenheim et. al., "Signal Analysis by Homomorphic Prediction," IEEE Trans., Acoust. Speech, Signal Processing, Vol. ASSP-24, No. 4, pp. 327-332, Aug. 1976.

[16] A. M. Noll, "Cepstrum pitch determination," J. Acoust., Soc. Am., Vol. 41, No. 2, pp. 293-309, Feb. 1967.

[17] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., Vol. 55, No. 6, pp. 1304-1312, June 1974.