JOINT OPTIMIZATION OF SHORT-TERM AND LONG-TERM PREDICTORS IN CELP SPEECH CODERS

Houman Zarrinkoub[1,2], *Paul Mermelstein*[1]
[1] Institut National de la Recherche Scientifique , Université du Québec, Montreal, Canada
[2] The MathWorks, Natick, MA, USA

ABSTRACT

The objective of this work is to investigate whether joint optimization of short-term and long-term predictors manifests significant advantages over the sequential optimization in speech coding. We propose a new joint optimization method based on Wiener filtering. The proposed analysis model resolves the pitch-bias problem of classical LPC analysis by considering the contribution of the long-term predictor while optimizing the short-term predictor. Our approach to joint optimization is based on analysis-by-synthesis and guarantees the synthesis filter stability. By applying our proposed joint optimization approach to CELP coding we obtain superior objective and subjective performance relative to CELP coding with sequential optimization. To provide voice quality equivalent to that of sequentially optimized CELP, the jointly optimized coder needs fewer FCB pulses and requires a reduced bit budget for LPC quantization. Our listening tests suggest that the JCELP coder at 4.25 kbps is equivalent in quality to the G.729 at 8 kbps.

1. INTRODUCTION

CELP coding has established itself as the dominant technology for voice compression in the past two decades. The CELP synthesis model is based on an auto-regressive (AR) source-filter representation. In each frame, the speech is synthesized as the output of an all-pole filter driven by an excitation signal. The excitation signal is composed of an adaptive component modeling the periodic, correlated and predictable part and a fixed component modeling the non-periodic, non-predictable and noisy part of the excitation.

The traditional approach to the optimization of the CELP synthesis model parameters is based on a sequential estimation. First, a classical LPC analysis optimizes the LPC filter and then based on an analysis by synthesis approach the adaptive and the fixed components of the excitation are optimized. Despite the longevity of this sequential approach, prior work has pointed out many of its limitations. These limitations become critical for speech coding at rates near 4 kbps, where the FCB pulse density and hence the contribution of the fixed component is severely limited.

Atal [1] and Makhoul [2] have separately studied the pitch-bias problem of the LPC analysis. Atal et al noted that the classical LPC analysis is optimal only if the input signal to an AR synthesis model is spectrally white. By ignoring the periodicity and the correlations of voiced speech, the LPC analysis becomes a sub-optimal approach. Makhoul et al have demonstrated that the classical LPC filters are optimized for a correlation function that is an aliased version of the true speech auto-correlation. Oudot et al. [3] noted that the peaks of the LPC spectra are artificially biased toward the pitch harmonics. Based on a harmonic spectral analysis, Murthi and Rao [4] provide a theoretical basis for these observations. By minimizing the LP residual power, the LPC analysis over-estimates the LPC filter gain. Hence, the poles of the LPC filter manifest a pattern of excessive resonance and the magnitude spectra exhibit sharper contours and larger dynamic ranges [1-4]. These spectral features necessitate a higher bit budget for quantization and more importantly degrade the performance of the ensuing long-term predictor. The sub-optimality of the sequential LP estimation was first studied by Kabal and Ramachandran [5] and later by Kabal and Zad-Issa [6]. They showed how the estimation noise introduced by the pitch-biased LPC analysis manifests itself as irregular variations of the pitch harmonics in the LPC residual signal [6]. This in turn affects the analysis-by-synthesis process, results in a reduced adaptive component contribution and lowers the overall prediction gain [5].

To overcome the limitations of the sequential LP estimation, various joint optimization methods have been proposed in the literature [1,5,7-9]. However, these methods suffer from many or all of these shortcomings:

1) The error minimization function will not fit an AR synthesis model and will be based on an auto-regressive-exogenous (ARX) model. As a result, the stability of the resulting LPC filter may not always be guaranteed.

2) The error minimization for both the LPC and the excitation parameters is specified solely in the LP residual domain. As a result of this choice for the optimization domain we lose all the advantages associated with the classical CELP sequential estimation, which optimizes the excitation signal in the perceptual synthesis domain.

3) The synthesis noise feedback and the advantages associated with analysis by synthesis are ignored.

Our approach to joint optimization addresses these shortcomings by formulating the design based on a Wiener filtering foundation and by integrating the analysis by synthesis into the overall error minimization.

This paper is organized as follows: In section 2, we first present the theory behind the proposed methodology and discuss the integration of our algorithm as part of a CELP coder. Section 3 will contain the main results showing the performance advantages of the joint optimization relative to the sequential optimization. Finally in section 4, we provide a summary and some concluding remarks.

0-7803-7663-3/03/$17.00 ©2003 IEEE

2. THEORY

Our approach to the joint optimization is based on analysis by synthesis and is composed of two steps: an open-loop step followed by a closed loop step. In the open loop step, we optimize the LPC filter given a viable candidate for the adaptive component. In the ensuing closed loop step, we simultaneously find the best filter and the best adaptive component candidate, by choosing the set that minimizes the error in the perceptual and synthesis domain.

In the open loop joint optimization step, the objective is to optimize the LPC prediction error filter $A(z)$ given a desired value for the filter output. This is a Wiener filtering problem as follows:

Let the speech signal $s(n)$ be the filter input, let the LPC residual signal $x(n)$ be the filter output and let the objective be to minimize the mean square error between the LPC residual and a desired signal $r(n)$.

$$e(n) = x(n) - r(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) - r(n)$$

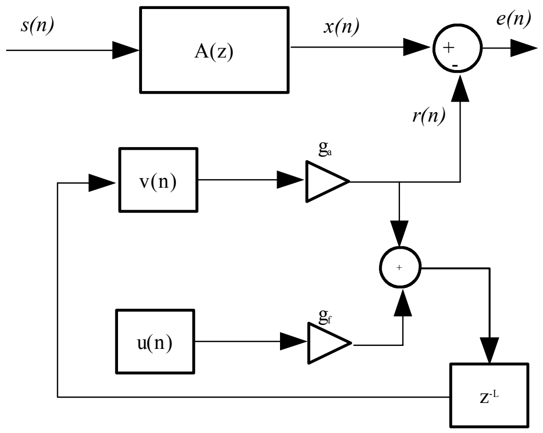

Figure 1 CELP-optimized Wiener filtering in the proposed open loop joint optimization method

The choice of the desired signal is the distinguishing feature of our proposed method. We define the adaptive component of the excitation as the desired signal, i.e., $r(n) = g_a v(n)$, where $v(n)$ represents the adaptive codebook signal and g_a represents the adaptive codebook gain. This choice results in a CELP-optimized joint estimation method. Since the desired signal is the adaptive component of the excitation, we are in fact approximating the residual signal to the best quantized long-term predictable signal that the CELP model can produce.

To minimize the distance measure, the error signal must be orthogonal to the space spanned by the linear prediction filter:

$$\langle e(n) | s(n-i) \rangle = \sum_{n=0}^{N-1} e(n)s(n-i) = 0 \quad i = 1,...,p$$

The solution for the Wiener filtering problem is expressed in terms of p simultaneous normal equations:

$$\sum_{j=1}^{p} a_j R_s(|i-j|) = R_s(i) + R_{sr}(i) \quad i = 1,...,p$$

Here, $R_s(i) = \sum_{n=0}^{N-1} s(n)s(n-i)$ is the auto-correlation sequence of the speech and $R_{sr}(i) = \sum_{n=0}^{N-1} r(n)s(n-i)$ is the cross-correlation between the speech and the desired signal. The normal equation is similar to that of the classical LPC analysis.

$$\boldsymbol{Ra} = \underline{\boldsymbol{p_C}} + \underline{\boldsymbol{p_J}}$$

The common terms are the auto-correlation matrix $\boldsymbol{R}$, the LPC solution vector $\boldsymbol{a}$ and the correlation vector $\underline{\boldsymbol{p_C}}$. Note that if the cross-correlation vector $\underline{\boldsymbol{p_J}}$ vanishes, i.e., the cross-correlation between the speech and the adaptive component is ignored, the joint prediction estimation reduces to the classical LPC analysis. Also note that the new short-term predictor solution may be expressed as the classical LPC solution with an extra correction term:

$$\underline{\boldsymbol{a}}_{\boldsymbol{N}} = \boldsymbol{R}^{-1}\underline{\boldsymbol{p}} = \boldsymbol{R}^{-1}(\underline{\boldsymbol{p_C}} + \underline{\boldsymbol{p_J}}) = \underline{\boldsymbol{a}}_{\boldsymbol{c}} + (\boldsymbol{R}^{-1}\underline{\boldsymbol{p_J}})$$

Since the Toeplitz auto-correlation matrices of the two analysis methods are identical, the JCELP method guarantees a stable synthesis filter. Note also that the computation of the jointly optimized LPC parameters does not require a matrix inversion for each viable candidate of the desired signal. This fact addresses the concern about the additional complexity associated with the proposed algorithm.

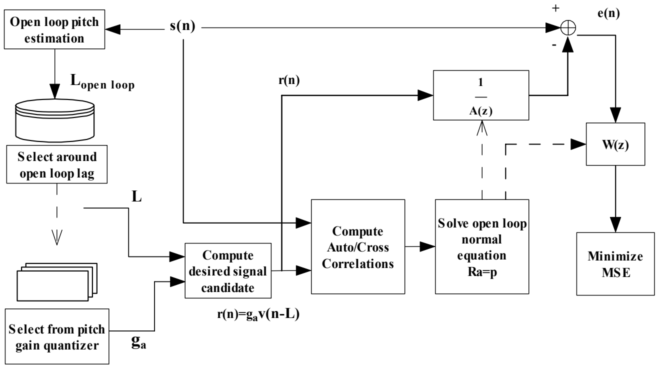
Open loop pitch estimation
s(n)
+
e(n)
-
L_open loop
r(n)
1/A(z)
Select around open loop lag
W(z)
L
Compute desired signal candidate
Compute Auto/Cross Correlations
Solve open loop normal equation Ra=p
Minimize MSE
r(n)=g_av(n-L)
Select from pitch gain quantizer
g_a

Figure 2 Closed loop analysis by synthesis in the proposed joint optimization

In the closed loop step, the perceptually weighted error between the input and the contribution of the ACB signal to the synthesis is minimized. Hence, we are effectively maximizing the contribution of predictable components of the CELP model to the synthesis process. As a result, the overall optimization criterion minimizes the energy of the FCB target signal.

Figure 2 illustrates all the steps involved in the proposed joint optimization. For each desired signal candidate, having computed the optimal LPC parameters in the open loop step, the closed loop search will proceed as follows: First the perceptual weighting filter $W(z) = \dfrac{A(z)}{A(z/\gamma 2)}$ is computed.

Then the synthesis step generates the synthetic speech signal candidate. Finally, by perceptual weighting filtering both the input and the synthetic speech and minimizing the error the best short term and long term prediction parameters are found.

Note that in the context of a generalized CELP coder (such as RCELP) our closed loop search method only iterates around pitch gain candidates and uses the pre-defined pitch contour to modify both the input and the excitation signal.

3. RESULTS

We have integrated the proposed joint optimization into a CELP coder. Two distinct realizations of the JCELP coder were studied. The first realization is based on the non-generalized CELP method and employs constant pitch lag values in each sub-frame. It operates at a rate of around 5 kbps. The second realization is based on an RCELP generalized analysis-by-synthesis method and employs a smoothly evolving pitch contour. This coder operates at a rate of near 4 kbps.

In each sub-frame, following pre-processing and a voice-unvoiced decision step, the open loop pitch estimation provides a range of pitch lag values to be used in the joint optimization process. In unvoiced frames, we revert to the sequential estimation due to its optimality. In voiced frames, based on our proposed joint optimization the LPC, pitch lag and pitch gain parameters minimizing the perceptually weighted synthesis error are then determined. Following the joint estimation, the target signal for the FCB component of the excitation is computed. Then the parameters of the fixed component of the excitation signal, i.e. the FCB pulse positions and signs and the FCB gain, are optimized through a classical analysis by synthesis approach.

Table 1 details the bit allocation of both the relaxed and the conventional JCELP coder. The design is based on a frame size of 20 ms with 4 sub-frames of 5 ms.

Bits per *20 ms frame*	*Conventional* *JCELP(5.1 kbps)*	*Relaxed* *JCELP (4.25 kbps)*
LPC	17	17
Pitch lag	[8,4,8,4] = 24	7
Pitch gain	3 x 4 = 12	3 x 4 = 12
FCB pulses	10 x 4 = 40	10 x 4 = 40
FCB gain	9	9
Total	102	85

Table 1 Bit allocation of proposed JCELP coders

We performed subjective and objective assessments to gauge the performance of the JCELP coder as compared to the G.729 coder. Although the objective measures are not the most definitive means to rate the quality of various coders, they play an important role is calibrating various speech quality correlates. Figure 3 shows the variations of the total segmental SNR as a function of the perceptual weighting filter shaping parameter. JCELP consistently outperforms the classical coder in all error minimization domains by a range of 0.5 to 2 dB. The results show that as we move from error minimization in the residual domain with a shaping factor of zero, to that in the perceptual and synthesis domains with shaping factors near one, the segmental SNR values of JCELP increasingly exceed those of the classical CELP. The results of our objective experiments were computed over a speech corpus containing 20 minutes of clean speech.

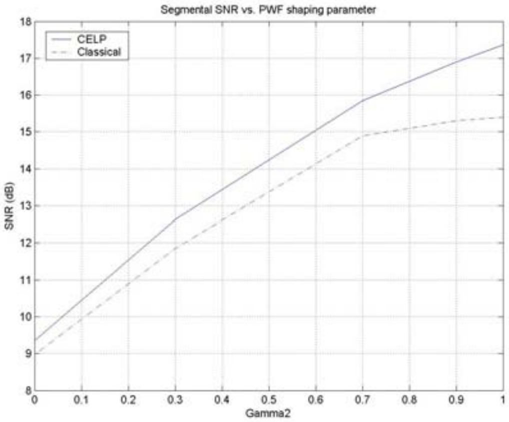

Figure 3 Segmental SNR & perceptual filter parameter

Figure 4 below shows the consistent superiority of the JCELP in terms of the total SNR measure relative to the classical CELP. The margin of improvement for the SNR ranges from an average of about 2 dB at low pulse densities to about 2.5 dB in high pulse densities. Also note that to achieve an average segmental SNR of about 15 dB, in classical CELP a density of 4 pulses per subframe is needed whereas in JCELP the same performance is attainable with about 2 pulses per subframe.

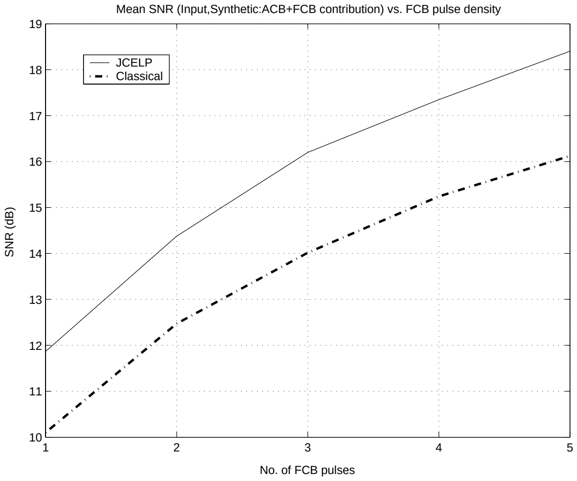

Figure 4 Segmental SNR as a function of pulse density

Figure 5 shows the mean magnitude response of the JCELP and classical CELP LPC filters. We can readily observe that in contrast to classical CELP, the JCELP spectrum exhibits a lower dynamic range, higher pole bandwidths and a lower filter gain. Hence, all the spectral artifacts of the pitch-biased LPC analysis are addressed. Note that a flatter LPC filter spectrum signifies a reduced variance for the LSF parameters representing the filter coefficients.

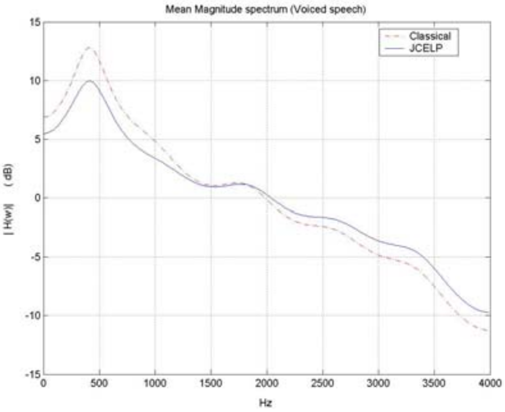

Figure 5 Mean spectra of JCELP and classical CELP

Table 2 showcases how the desirable spectral features of the jointly optimized LPC parameters translate into bit saving in parameter quantization. The results are based on profiling the mean values and the 2 dB outliers of the spectral distortion measure. We observe that transparent quantization is obtained by using an 18 bits per frame switched predictive split VQ quantizer. In an earlier work [10], we have shown that for transparent LSF quantization in a classical CELP coder using the same algorithm a 21-bit quantizer is needed.

Parameter type	Bit Allocation	S.D.	
		Mean (dB)	Outlier
Classical LSF	21(10,10,1)	1.13	3.6
	20(10,9,1)	1.20	5.1
Jointly optimized LSF	18(9,8,1)	1.04	2.1
	17(8,8,1)	1.10	3.8

Table 2 Spectral distortion profiles of LSF quantizers

To assess the subjective quality of the proposed coder, we performed an informal listening test. We used 32 different speech utterances, uttered by 4 male and 4 female speakers. Sixteen listeners, 7 female and 9 male, participated in the evaluation.

Type	*Input signal*	*G.711 64 kb/s*	*G.729 8 kb/s*	*JCELP 4.2 kb/s*	*G.723 5.3 kb/s*
Score	**4.15 ± 0.22**	**3.97 ± 0.19**	**3.35 ± 0.22**	**3.32 ± 0.23**	**2.19 ± 0.30**

Table 3 MOS scores from the subjective evaluation

Table 3 summarizes the overall MOS score results in terms of averages and their standard deviations. Based on these results, we can safely state that the G.729 at 8 kbps and JCELP at 4.25 kbps have equivalent voice qualities.

The added complexity of our joint optimization method is concentrated in the open loop step. For each viable candidate of the desired signal, the normal equation correction term has to be computed in each sub-frame. Based on our estimates, the results may be an additional 5% to the overall complexity of a typical CELP encoder.

4. SUMMARY

Joint optimization of short-term and long-term predictors in a CELP synthesis model based on Wiener filtering is advantageous for coding voiced speech. Its simple formulation provides significant advantages in terms of rate-distortion performance. It results in a synthesis filter requiring fewer bits for quantization and an excitation signal requiring fewer pulses, relative to the classical CELP coder to obtain the same voice quality. When implemented as part a 4.25 kbps RCELP coder, the proposed optimization enables the coder to provide equivalent quality to the ITU G.729 coder at 8 kbps.

5. REFERENCES

[1] S. Singhal, B. S. Atal, "Optimizing LPC parameters for multi-pulse excitation," *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, pp. 781-784, April 1983

[2] A. El-Jaroudi, J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. on Signal Processing*, vol. 39, no. 2, pp. 411-423, February 1991

[3] M.C. Oudot. O. Cappe, E. Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 469-481, July 2001

[4] M.N. Murthi, B.D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response system," *IEEE Trans. on Speech and Audio Pro*cessing, vol. 8, no. 3, pp. 221-239, May 2000

[5] P. Kabal, A. Ramachandran, "Joint optimization of linear predictors in speech coders," *IEEE Trans. on ASSP*, vol. 37, pp. no. 5, 642-650, May 1989

[6] M. Zad-Issa, P. Kabal, "A new LPC error criterion for improved pitch tracking," *IEEE Workshop on Speech Coding*, Pennsylvania, USA, pp. 1-3, September 1997

[7] Per Hedelin, "A glottal LPC-vocoder," *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, vol. 1, pp. 1.6.1-1.6.4, 1984

[8] K. Hacioglu, A. Hasib, "Pulse-by-pulse re-optimization of the synthesis filter in Pulse-based Coders," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, March 1998

[9] M. Fratti, G.A. Mian, G. Riccondi, "An approach to parameter re-optimization in multipulse-based coders," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 463-465, October 1993

[10] H. Zarrinkoub, P. Mermelstein, "Switched Prediction and Quantization of LSP Frequencies," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 757-760, Atlanta, GA, May 1996*