

Low Complexity Joint Optimization of Excitation Parameters in Analysis-by-Synthesis Speech Coding

U. Mittal, J. P. Ashley, E. M. Cruz-Zeno

Motorola Labs
1301 East Algonquin Road, Schaumburg, IL 60196, USA
mittal, ashley, cruz@labs.mot.com

Abstract

Codebook searches in analysis-by-synthesis speech coders typically involve minimization of a perceptually weighted squared error signal. Minimization of the error over multiple codebooks is often done in a sequential manner, resulting in the choice of overall excitation parameters being sub-optimal. In this paper, we propose a joint excitation parameter optimization framework in which the associated complexity is slightly greater than the traditional sequential optimization, but with significant quality improvement. Moreover, the framework allows joint optimization to be easily incorporated into existing pulse codebook systems with little or no impact to the codebook search algorithms.

1. Introduction

Algebraic Code Excited Linear Prediction (ACELP), which is used in many speech-coding standards, solves the inherent time complexity issues of a family of analysis-by-synthesis based Code Excited Linear Predictive (CELP) speech-coders. In typical CELP coders, the synthetic speech is obtained by passing a synthetic excitation vector through a linear prediction filter. The excitation vector is a weighted sum of an adaptive codebook (ACB) excitation and a fixed codebook (FCB) excitation. The search process finds the best excitation candidate vector from both codebooks and also computes their respective gains. The search in standard CELP coders is generally sequential, i.e., the parameters (gain and the code vector index) for the fixed codebook are obtained only after the parameters for adaptive codebook have been found and the contribution of the adaptive codebook has been subtracted from the weighted speech. Ideally, the parameters of both codebooks should be obtained jointly, but the computational complexity generally prohibits such an optimization.

In [1,2] lower complexity joint codebook optimizations have been proposed. Here, the codebook search methods start with a primary search to obtain the adaptive codebook parameters and then fix the adaptive codebook excitation vector (but not the gain) to obtain both adaptive and fixed codebook gains and the fixed codebook excitation. It was also shown in [1] that using such a joint optimization rather than a complete joint optimization (including the adaptive codebook excitation) does not result in significant loss of speech quality. The drawback of the method proposed in [1] is that it is at least 30% more complex than a similar sequential optimization. Moreover, it cannot be readily incorporated into existing FCB search techniques.

In this paper, we propose a very low complexity joint FCB/ACB gains and FCB excitation optimization by modification of the correlation matrix used in the standard

FCB search methods. The modification of the correlation matrix enables use of standard FCB search algorithms for the joint optimization.

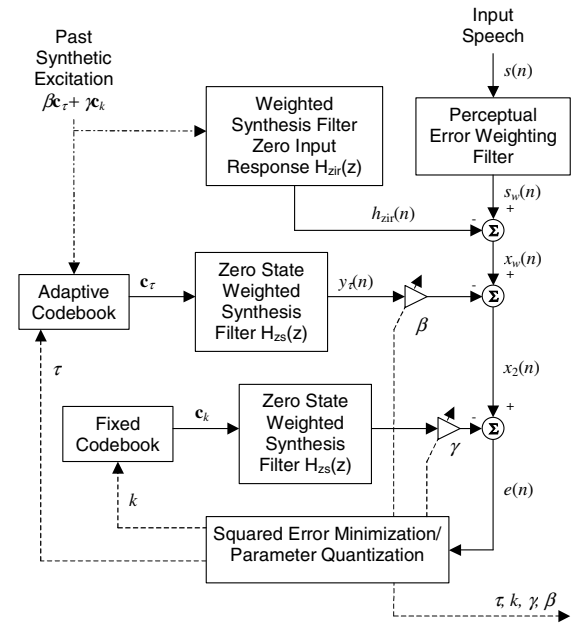


Figure 1: Typical Analysis-by-Synthesis Encoder

2. Joint Optimization

We use the notations from Figure 1 and Figure 2 and represent their column vector forms in bold face. The weighted error $e(n)$ in Figure 1 is given by:

$$\mathbf{e} = \mathbf{x}_w - \beta \mathbf{H} \mathbf{c}_\tau - \gamma \mathbf{H} \mathbf{c}_k \quad (1)$$

where \mathbf{H} is the matrix representing the weighted synthesis filters impulse response. The square error is given by:

$$\mathcal{E} = \|\mathbf{e}\|^2 = \|\mathbf{x}_w - \beta \mathbf{H} \mathbf{c}_\tau - \gamma \mathbf{H} \mathbf{c}_k\|^2 \quad (2)$$

In the typical sequential optimization, we first perform the ACB search by setting $\gamma = 0$ to obtain

$$\tau^* = \arg \max_{\tau} \left\{ \begin{array}{l} (\mathbf{x}_w^T \mathbf{H} \mathbf{c}_\tau)^2 \\ \mathbf{c}_\tau^T \mathbf{H}^T \mathbf{H} \mathbf{c}_\tau \end{array} \right\} \quad (3)$$

where τ^* is the value of τ that maximizes the bracketed expression, and

$$\beta = \frac{\mathbf{x}_w^T \mathbf{y}_\tau}{\mathbf{y}_\tau^T \mathbf{y}_\tau} \quad (4)$$

where $\mathbf{y}_\tau = \mathbf{H}\mathbf{c}_\tau$ is the filtered ACB excitation. Similarly, for the FCB search:

$$\mathcal{E} = \|\mathbf{x}_w - \beta\mathbf{H}\mathbf{c}_\tau - \gamma\mathbf{H}\mathbf{c}_k\|^2 = \|\mathbf{x}_2 - \gamma\mathbf{H}\mathbf{c}_k\|^2 \quad (5)$$

$$k^* = \arg \max_k \left\{ \frac{(\mathbf{d}_2^T \mathbf{c}_k)^2}{\mathbf{c}_k^T \Phi \mathbf{c}_k} \right\} \quad (6)$$

where $\mathbf{d}_2^T = \mathbf{x}_2^T \mathbf{H}$ and $\Phi = \mathbf{H}^T \mathbf{H}$.

Let us look at the joint optimization now. Going back to Eq. 2 and equating the partial differentials with respect to β and γ to zero results in

$$\mathbf{x}_w^T \mathbf{H} [\mathbf{c}_\tau \quad \mathbf{c}_k] = [\beta \quad \gamma] \begin{bmatrix} \mathbf{c}_\tau^T \mathbf{H}^T \mathbf{H} \mathbf{c}_\tau & \mathbf{c}_k^T \mathbf{H}^T \mathbf{H} \mathbf{c}_\tau \\ \mathbf{c}_\tau^T \mathbf{H}^T \mathbf{H} \mathbf{c}_k & \mathbf{c}_k^T \mathbf{H}^T \mathbf{H} \mathbf{c}_k \end{bmatrix}. \quad (7)$$

Letting $\mathbf{d}^T = \mathbf{x}_w^T \mathbf{H}$, $\Phi = \mathbf{H}^T \mathbf{H}$, and $\mathbf{C} = [\mathbf{c}_\tau \quad \mathbf{c}_k]$, Eq. 7 can be written as

$$\mathbf{d}^T \mathbf{C} = [\beta \quad \gamma] \mathbf{C}^T \Phi \mathbf{C}. \quad (8)$$

From Eq. 8, letting $\mathbf{g} = [\beta \quad \gamma]$ we get

$$\mathbf{g} = \mathbf{d}^T \mathbf{C} [\mathbf{C}^T \Phi \mathbf{C}]^{-1}. \quad (9)$$

Applying \mathbf{g} to Eq. 2 we get:

$$\mathcal{E} = \mathbf{x}_w^T \mathbf{x}_w - \mathbf{d}^T \mathbf{C} [\mathbf{C}^T \Phi \mathbf{C}]^{-1} \mathbf{C}^T \mathbf{d}. \quad (10)$$

Thus, to minimize \mathcal{E} ,

$$[\tau^* \quad k^*] = \arg \max_{\tau, k} \left\{ \mathbf{d}^T \mathbf{C} [\mathbf{C}^T \Phi \mathbf{C}]^{-1} \mathbf{C}^T \mathbf{d} \right\}. \quad (11)$$

In a strict sense, this represents the simultaneously joint optimization of both ACB and FCB codevectors, and their associated gains. In practice, however, this joint optimization is prohibitively complex. As a simplified alternative, we wish to assume that the ACB codevector \mathbf{c}_τ is determined *a priori* (via Eq. 3), and the remaining parameters \mathbf{c}_k , β , and γ are determined in a jointly optimal fashion. So, moving back to Eq. 11, we can begin by expanding and eliminating terms that are independent of \mathbf{c}_k . By inverting the inner matrix $\mathbf{C}^T \Phi \mathbf{C}$ and substituting temporary variables in Eq. 11 we get:

$$k^* = \arg \max_k \left\{ \frac{1}{D_k} (MA_k^2 - 2NA_k B_k + R_k N^2) \right\} \quad (12)$$

where $M = \mathbf{c}_\tau^T \Phi \mathbf{c}_\tau$, $N = \mathbf{d}^T \mathbf{c}_\tau$, $B_k = \mathbf{c}_\tau^T \Phi \mathbf{c}_k$, $A_k = \mathbf{d}^T \mathbf{c}_k$, and $R_k = \mathbf{c}_k^T \Phi \mathbf{c}_k$. Also, $D_k = \mathbf{c}_\tau^T \Phi \mathbf{c}_\tau \mathbf{c}_k^T \Phi \mathbf{c}_k - \mathbf{c}_\tau^T \Phi \mathbf{c}_\tau \mathbf{c}_\tau^T \Phi \mathbf{c}_k = MR_k - B_k^2$ is the determinant of the inverted matrix. Note that M is the energy of the filtered ACB excitation vector, N is the correlation between weighted speech and filtered ACB excitation, A_k is the correlation between the reverse filtered target vector and FCB excitation, and B_k is the correlation between filtered ACB excitation and filtered FCB excitation. Since M and N^2 are both non-negative and are independent of k , instead of solving Eq. 12 we can equivalently solve:

$$k^* = \arg \max_k \left\{ \frac{M}{N^2 D_k} (MA_k^2 - 2NA_k B_k + R_k N^2) \right\}. \quad (13)$$

If we define $a_k = MA_k$, $b_k = NB_k$, $R'_k = MN^2 R_k$, $D'_k = N^2 D_k$, we get:

$$k^* = \arg \max_k \left\{ \frac{1}{D'_k} (a_k^2 - 2a_k b_k + R'_k) \right\}. \quad (14)$$

We can now express R'_k in terms of D'_k by observing that since $D'_k = N^2 D_k = N^2 MR_k - N^2 B_k^2$, $R'_k = MN^2 R_k$, and $b_k = NB_k$, then $R'_k = D'_k + b_k^2$. Substituting into Eq. 14 yields the following:

$$k^* = \arg \max_k \left\{ \frac{1}{D'_k} (a_k^2 - 2a_k b_k + D'_k + b_k^2) \right\} \quad (15a)$$

$$k^* = \arg \max_k \left\{ \frac{1}{D'_k} ((a_k - b_k)^2 + D'_k) \right\} \quad (15b)$$

$$k^* = \arg \max_k \left\{ \frac{(a_k - b_k)^2}{D'_k} + 1 \right\}. \quad (15c)$$

Since the constant in Eq 15c has no effect on the maximization process, it can be removed, leaving:

$$k^* = \arg \max_k \left\{ \frac{(a_k - b_k)^2}{D'_k} \right\}. \quad (16)$$

Now, we will show that the parameters of the joint optimization can be transformed to the two pre-computed parameters of the sequential FCB optimization thereby enabling a sequential FCB search algorithm to be used for joint optimization. The two pre-computed parameters are the correlation matrix and the reverse filtered weighted target signal. Consider the sequential search based CELP coders. Referring back to Eq. 6, the numerator is the square of the dot product of FCB vector and a vector independent of k , and the denominator in a form $\mathbf{c}_k^T \Phi \mathbf{c}_k$, where Φ is a matrix that is also independent of k . We will now modify Eq. 16 so that it can also be written in the same form as Eq. 6. So if we first define the numerator in Eq. 16 to be "like" the numerator in Eq. 6 and equate like terms, we get:

$$\mathbf{d}^T \mathbf{c}_k \Leftrightarrow a_k - b_k \quad (17)$$

$$\mathbf{d}^T \mathbf{c}_k \Leftrightarrow MA_k - NB_k \quad (17a)$$

$$\mathbf{d}^T \mathbf{c}_k \Leftrightarrow (\mathbf{c}_\tau^T \Phi \mathbf{c}_\tau) \mathbf{d}^T \mathbf{c}_k - (\mathbf{d}^T \mathbf{c}_\tau) \mathbf{c}_\tau^T \Phi \mathbf{c}_k \quad (17b)$$

$$\mathbf{d}^T \mathbf{c}_k \Leftrightarrow (\mathbf{y}_\tau^T \mathbf{y}_\tau) \mathbf{x}_w^T \mathbf{H} \mathbf{c}_k - (\mathbf{x}_w^T \mathbf{y}_\tau) \mathbf{y}_\tau^T \mathbf{H} \mathbf{c}_k \quad (17c)$$

$$\mathbf{d}^T = ((\mathbf{y}_\tau^T \mathbf{y}_\tau) \mathbf{x}_w^T - (\mathbf{x}_w^T \mathbf{y}_\tau) \mathbf{y}_\tau^T) \mathbf{H}. \quad (18)$$

From this equation it can be seen that if the optimal ACB gain for the sequential search were used (from Eq. 4), and also noting (from Eq. 5) that $\mathbf{d}_2^T = \mathbf{x}_2^T \mathbf{H} = (\mathbf{x}_w - \beta \mathbf{y}_\tau)^T \mathbf{H}$, we can infer

$$\mathbf{d}^T = (\mathbf{y}_\tau^T \mathbf{y}_\tau) \mathbf{d}_2^T. \quad (19)$$

This says that the numerator of Eq. 16 is merely a scaled version of the numerator in Eq. 6.

Now moving to the denominator portion of Eq. 16. As in the numerator discussion above, we would now like to put the denominator in a form that is similar to that of Eq. 6. That is:

$$\mathbf{c}_k^T \Phi \mathbf{c}_k \Leftrightarrow D'_k \quad (20)$$

By using substitution of previously defined terms, we can derive the following sequence of equivalent expressions:

$$\mathbf{c}_k^T \Phi \mathbf{c}_k \Leftrightarrow N^2 MR_k - N^2 B_k^2 \quad (20a)$$

$$\mathbf{c}_k^T \Phi' \mathbf{c}_k \Leftrightarrow N^2 \mathbf{M} \mathbf{c}_k^T \Phi \mathbf{c}_k - N^2 (\mathbf{c}_\tau^T \Phi \mathbf{c}_k)^2 \quad (20b)$$

Since $\Phi = \mathbf{H}^T \mathbf{H}$ is symmetric, $\Phi = \Phi^T = \mathbf{H}^T \mathbf{H}$:

$$\mathbf{c}_k^T \Phi' \mathbf{c}_k \Leftrightarrow N^2 \mathbf{M} \mathbf{c}_k^T \Phi \mathbf{c}_k - N^2 \mathbf{c}_k^T \Phi \mathbf{c}_\tau \mathbf{c}_\tau^T \Phi \mathbf{c}_k \quad (20c)$$

$$\mathbf{c}_k^T \Phi' \mathbf{c}_k \Leftrightarrow \mathbf{c}_k^T (N^2 \mathbf{M} \Phi - N^2 \Phi \mathbf{c}_\tau \mathbf{c}_\tau^T \Phi) \mathbf{c}_k \quad (20d)$$

$$\mathbf{c}_k^T \Phi' \mathbf{c}_k \Leftrightarrow \mathbf{c}_k^T (N^2 \mathbf{M} \Phi - N^2 \mathbf{H}^T \mathbf{y}_\tau \mathbf{y}_\tau^T \mathbf{H}) \mathbf{c}_k \quad (20e)$$

Now letting $\mathbf{y} = \mathbf{H}^T \mathbf{y}_\tau$, we can rewrite Eq. 20e as:

$$\mathbf{c}_k^T \Phi' \mathbf{c}_k \Leftrightarrow \mathbf{c}_k^T (N^2 \mathbf{M} \Phi - N^2 \mathbf{y} \mathbf{y}^T) \mathbf{c}_k \quad (20f)$$

$$\Phi' = N^2 \mathbf{M} \Phi - N^2 \mathbf{y} \mathbf{y}^T \quad (21)$$

Therefore, the Eq. 16 can be written as:

$$k^* = \arg \max_k \left\{ \frac{(\mathbf{d}'^T \mathbf{c}_k)^2}{\mathbf{c}_k^T \Phi' \mathbf{c}_k} \right\} \quad (22)$$

$$k^* = \arg \max_k \left\{ \frac{(\mathbf{M} \mathbf{d}'^T \mathbf{c}_k)^2}{\mathbf{c}_k^T (N^2 \mathbf{M} \Phi - N^2 \mathbf{y} \mathbf{y}^T) \mathbf{c}_k} \right\}. \quad (23)$$

This shows that since the form of Eqs. 6 and 23 are generally the same, the terms \mathbf{d}' and Φ' can be pre-computed, and any existing sequential optimization algorithm may be transformed to a joint optimization without modification. Going back to Eq. 23, if the vector $\mathbf{y} = \mathbf{0}$, then the expression for the joint search would be equivalent to the corresponding expression for the sequential search. This implies that we can easily adaptively choose to do sequential search whenever needed.

2.1. Pitch Inclusion in Joint Optimization

In the above derivation the weighted synthesis filter \mathbf{H} for ACB excitation and FCB excitation were the same. But in subframes where the pitch delay is less than the subframe length, then a zero state pitch filter is generally included in the path of FCB excitation as shown in Figure 2. The pitch filter is given by

$$P(z) = \frac{1}{1 - \beta' z^{-\tau}}. \quad (24)$$

This makes the weighted synthesis filter different for the two excitations. Let \mathbf{H}_1 be the matrix representation of the weighted synthesis filter for ACB excitation, and \mathbf{H} be the matrix representation for the weighted synthesis filter for the FCB excitation, including pitch filtering, i.e., $\mathbf{H} = \mathbf{P} \mathbf{H}_1$. Now

$$\mathbf{e} = \mathbf{x}_w - \beta \mathbf{H}_1 \mathbf{c}_\tau - \gamma \mathbf{H} \mathbf{c}_k. \quad (25)$$

Proceeding as before we will again get Eq. 23 as the solution to the optimization problem. Note that now

$$\mathbf{y}_\tau = \mathbf{H}_1 \mathbf{c}_\tau. \quad (26)$$

In the sequential optimization approach, where the ACB gain is quantized before searching the FCB excitation, the pitch pre-filter coefficient β' can be derived from the quantized ACB gain. This allows the encoder and decoder to use consistent values of β' . In joint optimization, the ACB gain is obtained during the FCB excitation search, hence we cannot obtain β' from the quantized ACB gain. This problem also arises in speech coders having sequential optimization and then performing vector quantization of ACB and FCB gains [3,4]. Here β' is either chosen as a known constant [3] or is derived from the quantized ACB gain of the previous subframe [4].

In our approach, we obtain an initial estimate of β' from the ACB gain associated with the ACB search given in Eqs. 3 and 4. Next, the joint optimization according to Eq. 23 is performed (using $\mathbf{H} = \mathbf{P} \mathbf{H}_1$) resulting in the FCB shape vector \mathbf{c}_k . Finally, a 3-dimensional gain vector quantizer (for β' , β , and γ) is employed to minimize the squared error of Eq. 25, thus making the final version of β' consistent at both the encoder and decoder.

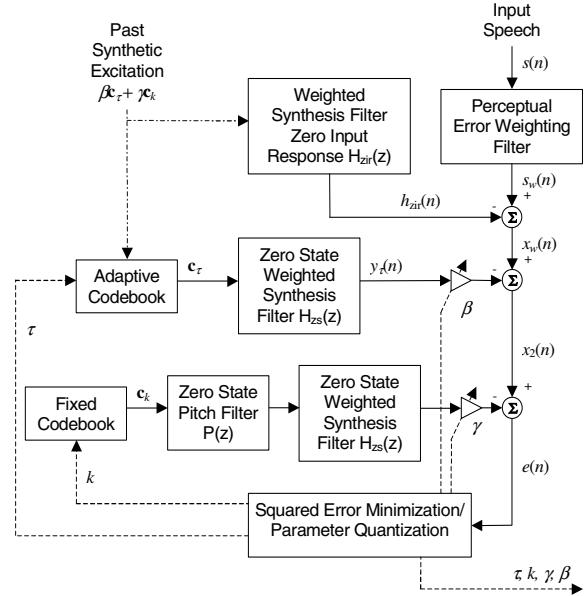


Figure 2: Analysis-by-Synthesis Encoder with Pitch Filtering

2.2. Complexity

Let us look at the numerator of the joint optimization (Eq. 23) and compare it to the numerator of sequential FCB optimization (Eq. 6). Note that for a given subframe length L , the additional complexity is L multiplies per subframe to get the numerator. Since $\mathbf{M} = \mathbf{y}_\tau^T \mathbf{y}_\tau$ already exists for the computation of the optimal τ in Eq. 3, no additional computations are necessary. The same is true for the computation of $\mathbf{N} = \mathbf{x}_w^T \mathbf{y}_\tau$. For the denominator, the generation of $\mathbf{y} = \mathbf{H}^T \mathbf{y}_\tau$ requires approximately one half of a length L linear convolution, or about $L(L+1)/2$ multiply-accumulate (MAC) operations. The $N^2 M$ scaling of the correlation matrix Φ can be efficiently implemented by scaling the elements of the impulse response $h(n)$ by $\sqrt{N^2 M}$ prior to generation of the matrix $\Phi = \mathbf{H}^T \mathbf{H}$. This requires a square root operation and about L multiplications. Similarly, scaling of the \mathbf{y} vector by N requires only about L multiply operations. Lastly, the generation and subtraction of the scaled $\mathbf{y} \mathbf{y}^T$ matrix from the scaled Φ , a $L \times L$ matrix, requires about $L(L+1)/2$

MAC operations. This is because $\mathbf{Y} = \mathbf{y}\mathbf{y}^T$ is a rank one matrix (i.e., $Y(i, j) = y(i)y(j)$) and now Φ' (removing scaling constants) can be generated as:

$$\phi'(i, j) = \phi(i, j) - y(i)y(j), \quad 0 \leq i < L, \quad 0 \leq j \leq i \quad (27)$$

As one may notice, only the upper or lower triangular part of the Φ' matrix needs to be generated because of symmetry. Thus, the total additional complexity required for a sequential to joint search transformation is about $L^2 + 4L$ multiply-accumulate operations per subframe. For a narrow band speech coder, typically $L=40$ which means around 1760 extra operations per 5 ms subframe. This is around 7% of the typical FCB search complexity, which is presumed to be around 5M operations/sec. For $L=70$ (wideband coder), 5200 extra operations per 5 ms subframe are needed, which is around 10% of a typical wideband coder's FCB search complexity (presuming 10M operations/sec). Thus, the complexity is considerably less than the method proposed in [1].

3. Results

For comparing the performances, we use two multi-rate coders in which silence is coded at 1.0 kbps and speech active frames are coded at 13.3 kbps and 15.9 kbps, respectively. Either the proposed joint search or the sequential search is used in the frames, which were coded at 13.3 kbps and 15.9 kbps. For both these coding rates, an *unconstrained* pulse codebook[5][6] is used for the fixed codebook. The number of pulses in a subframe of size 70 at 13.3 kbps and 15.9 kbps rates are 8 and 12, respectively. The 13.3 kbps rate was used as one of the coding rates of the Motorola's wideband speech coder submitted as a candidate for 3GPP2 wideband coder standardization activity. Since the proposed joint search and the sequential search minimize the energy of the weighted synthesis error, we use weighted signal to noise ratio (WSNR) and Avg-WSNR for comparison purposes. WSNR is the ratio of energy of weighted speech in speech active frames to the energy of weighted synthesis error in these frames. Avg-WSNR is the average of the ratio (in dB) of the energy of weighted speech of a frame to the energy of the weighted synthesis error of that frame. Again, this average is calculated over speech active frames. Table 1 shows the comparison of sequential to joint optimization for the 13.3 kbps coder, while Table 2 shows the comparison for the 15.9 kbps coder. The results shown in the tables are for 220 sec wideband speech having alternating sentence-pairs in male and female voices. The results clearly indicate that the joint search has a gain of around 0.15 dB in WSNR and 0.3 dB in Avg-WSNR.

Table 1: Performance of the 13.3 kbps Wideband Speech Coder

Optimization	WSNR (dB)	Avg-WSNR (dB)
Sequential	7.276	8.503
Joint	7.416	8.763

Table 2: Performance of the 15.9 kbps Wideband Speech Coder

Optimization	WSNR (dB)	Avg-WSNR (dB)
Sequential	8.366	9.540
Joint	8.527	9.837

4. Conclusions

A low complexity joint ACB/FCB gain and FCB excitation optimization method has been proposed. This method has complexity advantages over prior methods, and incurs only a 5% to 10% increase in complexity over similar sequential methods. Furthermore, this method can be easily incorporated into existing ACELP type speech coders through a simple translation of the correlation matrix. The search algorithm can remain unchanged.

This method has been shown to improve weighted SNR over known sequential methods by as much as 0.3 dB or the equivalent of about 650 bps in a 13.3 kbps wideband speech coder.

5. Acknowledgment

The authors would like to thank Mark Jasiuk of Motorola Labs' Speech Processing Research Lab for his technical contributions in handling the pitch filtering gain during joint optimization of the excitation parameters.

6. References

- [1] Woodard, J. P. and Hanzo, L., "Improvement of analysis-by-synthesis loop in CELP codecs," *IEEE Radio Receivers and Associated Systems Conference*, 114-118, 1995.
- [2] Gerson, I. A. and Jasiuk, M. A., "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps," *Proc. ICASSP*, 461-464, 1990
- [3] Salami, R., Laflamme, C., Adoul, J.-P., Massaloux, D., "A toll quality 8 Kb/s speech codec for personal communications system," *IEEE Tran. on Vehicular Technology*, 808-816, Aug. 1994.
- [4] Salami, R., Laflamme, C., Adoul, J.-P., Kataoba, A., Hayasi, S., Lamblin, C., Massaloux, D., Proust, S., Kroon, P., and Shoham, Y., "Description of the proposed ITU-T 8Kb/S speech coding standard," *IEEE Workshop on Speech Coding*, 3-4, 1995.
- [5] Ashley, J. P., Cruz-Zeno, E. M., Mittal, U., Peng, W., "Wideband coding of speech using a scalable pulse codebook," *IEEE Workshop on Speech Coding*, 148-150, 2000.
- [6] Mittal, U., Ashley, J. P., Cruz-Zeno, E. M., "Coding unconstrained FCB excitation using combinatorial and Huffman codes," *IEEE Workshop on Speech Coding*, 129-131, 2002.