JOINT OPTIMIZATION OF EXCITATION PARAMETERS IN ANALYSIS-BY-SYNTHESIS SPEECH CODERS HAVING MULTI-TAP LONG TERM PREDICTOR

Udar Mittal, James P. Ashley, Edgardo M. Cruz-Zeno, and Mark A. Jasiuk

Motorola Labs, 1301 East Algonquin Road, Schaumburg, IL 60196, USA
emails: mittal, ashley, cruz, jasiuk@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. This technique is part of the 3GPP2 Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB) Rate Set 1 Standard.

1. INTRODUCTION

Algebraic Code Excited Linear Prediction (ACELP) [3], 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 the gain scaled sum of an adaptive codebook (ACB) excitation and a fixed codebook (FCB) excitation. The gain associated with ACB excitation and FCB excitation are called ACB gain and FCB gain, respectively. The ACB excitation is obtained from the past excitations using a delay parameter which may have sub sample (fractional) resolution. As described in [4], a multi-tap filter applied to the sub-sample resolution ACB excitation vector frees the modeling of delay from the multi-tap filter and hence provides better spectral shaping to the ACB excitation. One can also view ACB excitation generated using the multi-tap filter approach as ACB excitation being generated as a weighted sum of more than one ACB vectors. In the subsequence, we will be using terms like multi-tap filter weights or ACB gains to describe these weights.

The FCB excitation is obtained from searching a stochastic codebook. The sequential search process first finds the best excitation candidate vector from the adaptive codebook and computes multi-tap filter taps. 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, all the parameters (ACB as well as FCB) should be obtained jointly, but the computational complexity generally prohibits such an optimization.

In [1,2] lower complexity joint codebook optimizations have been proposed. In these approaches, 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, and it cannot be readily incorporated into existing FCB search techniques. Furthermore, it cannot be easily extended to include the weights of the multi-tap filter when the filter order is more than one.

In this paper, we propose a very low complexity joint FCB excitation, FCB gain, and multi-tap filter weights optimization by modification of the correlation matrix used in the standard sequential FCB search methods. The modification of the correlation matrix enables use of standard FCB search algorithms for the joint optimization.

2. JOINT OPTIMIZATION

Let L be the subframe length, n ($n < L$) be the multi-tap filter order, β_i, $1 \le i \le n$ be the filter weights, and $\mathbf{c}_\tau^i, 1 \le i \le n$ be the ACB vectors. Let $\mathbf{H}_1$ and $\mathbf{H}_2$ be the matrix representation for the weighted synthesis filters for

ACB excitation and FCB excitation, respectively. Note that the weighted synthesis filter for ACB excitation and FCB excitation are considered to be different as the weighted synthesis filter for the FCB excitation may also include a pitch sharpening filter. The ACB excitation $\mathbf{c}_\tau$ is given by:

$$\mathbf{c}_\tau = \sum_{i=1}^{n} \beta_i \mathbf{c}_\tau^i \tag{1}$$

and the weighted synthetic contribution from the ACB is given by:

$$\mathbf{y}_\tau = \sum_{i=1}^{n} \beta_i \mathbf{y}_\tau^i \tag{2}$$

where $\mathbf{y}_\tau^i = \mathbf{H}_1 \mathbf{c}_\tau^i$.

Let $\mathbf{c}_k$ be the FCB excitation and γ the FCB gain. Define $\mathbf{y}_k = \mathbf{H}_2 \mathbf{c}_k$, $\mathbf{Y}_\tau = \left(\mathbf{y}_\tau^1 \quad \mathbf{y}_\tau^2 \quad \ldots \quad \mathbf{y}_\tau^n\right)$, $\mathbf{Y} = \left(\mathbf{y}_k \quad \mathbf{Y}_\tau\right)$, $\mathbf{B} = \left(\beta_1 \quad \beta_2 \quad \ldots \quad \beta_n\right)$, and $\mathbf{\Gamma} = \left(\gamma \quad \mathbf{B}\right)$.

Let $\mathbf{x}_w$ be the weighted target after the zero input response of the weighted synthetic filter has been removed from the weighted speech. The squared error, which needs to be minimized, is now given by:

$$\varepsilon = \left\|\mathbf{x}_w - \mathbf{y}_\tau - \gamma \mathbf{H}_2 \mathbf{c}_k\right\|^2 = \left\|\mathbf{x}_w - \mathbf{Y}_\tau \mathbf{B}' - \gamma \mathbf{y}_k\right\|^2 \tag{3}$$

where $\mathbf{B}'$ is the transpose of $\mathbf{B}$. In the typical sequential optimization, $\mathbf{c}_\tau^i$ and β_i are first obtained to minimize ε assuming $\gamma = 0$. It can be derived from Eq. 3 that

$$\mathbf{B}' = (\mathbf{Y}_\tau' \mathbf{Y}_\tau)^{-1} \mathbf{Y}_\tau' \mathbf{x}_w \tag{4}$$

Now for the FCB search

$$\varepsilon = \left\|\mathbf{x}_w - \mathbf{Y}_\tau \mathbf{B}' - \gamma \mathbf{y}_k\right\|^2 = \left\|\mathbf{x}_2 - \gamma \mathbf{H}_2 \mathbf{c}_k\right\|^2, \tag{5}$$

where $\mathbf{x}_2 = \mathbf{x}_w - \mathbf{Y}_\tau \mathbf{B}'$. The optimal FCB excitation is obtained using:

$$k^* = \underset{k}{\operatorname{argmax}} \left\{ \frac{\left(\mathbf{d}' \mathbf{c}_k\right)^2}{\mathbf{c}_k' \mathbf{\Phi} \mathbf{c}_k} \right\} \tag{6}$$

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

Let us look at the joint optimization now. Rewriting equation (3), we get:

$$\varepsilon = \left\|\mathbf{x}_w - \mathbf{Y} \mathbf{\Gamma}'\right\|^2 \tag{7}$$

From Eq. 7, to minimize ε,

$$\mathbf{\Gamma}' = (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}' \mathbf{x}_w \tag{8}$$

Substituting this value of Γ in (7), results in

$$\varepsilon = \mathbf{x}_w' \mathbf{x}_w - \mathbf{x}_w' \mathbf{Y} (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}' \mathbf{x}_w \tag{9}$$

Thus minimizing ε is equivalent to maximizing

$$\mathbf{x}_w' \mathbf{Y} (\mathbf{Y}'\mathbf{Y})^{-1} \mathbf{Y}' \mathbf{x}_w \tag{10}$$

Define $\mathbf{R} = \mathbf{Y}'\mathbf{Y}$, $\mathbf{R}_\tau = \mathbf{Y}_\tau' \mathbf{Y}_\tau$, $\mathbf{r} = \mathbf{Y}' \mathbf{x}_w$, and $\mathbf{r}_\tau = \mathbf{Y}_\tau' \mathbf{x}_w$. Also define $R_k = \mathbf{y}_k' \mathbf{y}_k = \mathbf{c}_k \mathbf{\Phi} \mathbf{c}_k$, $\mathbf{u} = \mathbf{Y}_\tau' \mathbf{y}_k$. Now Eq. 10 can be written as $\mathbf{r}' \mathbf{R}^{-1} \mathbf{r}$. In a strict sense, this represents the composite joint optimization of ACB and FCB excitation vectors and their gains. In practice, however, this joint optimization is prohibitively complex. As a simplified alternative, we assume that the ACB excitation vectors $\mathbf{c}_\tau^i$ are determined *a priori*, and the remaining parameters $\mathbf{c}_k$, β_i, and γ are determined in a jointly optimal fashion.

So, moving back to Eq. 10, we begin by expanding and eliminating terms that are independent of $\mathbf{c}_k$. We start by inverting the inner matrix $\mathbf{R} = \mathbf{Y}'\mathbf{Y}$. It can be verified that the inverse of $\mathbf{R}$ is given by

$$\mathbf{R}^{-1} = \begin{pmatrix} a & \mathbf{b}' \\ \mathbf{b} & \mathbf{Q} \end{pmatrix} \tag{11}$$

where $a = 1/(R_k - \mathbf{u}' \mathbf{R}_\tau^{-1} \mathbf{u})$, $\mathbf{b} = -a \mathbf{R}_\tau^{-1} \mathbf{u}$, and $\mathbf{Q} = \mathbf{R}_\tau^{-1} + a \mathbf{R}_\tau^{-1} \mathbf{u}\mathbf{u}' \mathbf{R}_\tau^{-1}$. Expanding Eq. 10, we get

$$a(\mathbf{x}_w' \mathbf{y}_k)^2 - 2a \mathbf{x}_w' \mathbf{y}_k \mathbf{u}' \mathbf{R}_\tau^{-1} \mathbf{u} + \mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{r}_\tau + a \mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{u}\mathbf{u}' \mathbf{R}_\tau^{-1} \mathbf{r}_\tau, \tag{12}$$

which needs to be maximized to get $\mathbf{c}_k$.

Since $\mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{r}_\tau$ is dependent only on ACB excitation vectors, the synthesis filter and the weighted target, the term is constant during the optimization process. Thus, the above term can be removed from Eq. 12 to get

$$k^* = \underset{k}{\operatorname{argmax}} \left\{ a(\mathbf{x}_w' \mathbf{y}_k - \mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{u})^2 \right\}, \tag{13}$$

$$= \underset{k}{\operatorname{argmax}} \left\{ \frac{(\mathbf{x}_w' \mathbf{y}_k - \mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{u})^2}{R_k - \mathbf{u}' \mathbf{R}_\tau^{-1} \mathbf{u}} \right\} \tag{13a}$$

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 $\mathbf{c}_k^T \mathbf{\Phi} \mathbf{c}_k$, where $\mathbf{\Phi}$ is a matrix that is also independent of *k*. We will now modify Eq. 13 so that it can also be written in the same form as Eq. 6.

Let us look as the numerator of Eq. 13a.

$$(\mathbf{x}_w' \mathbf{y}_k - \mathbf{r}_\tau' \mathbf{R}_\tau^{-1} \mathbf{u})^2 \tag{14}$$

$$= (\mathbf{x}_w' \mathbf{y}_k - \mathbf{B}\mathbf{Y}_\tau' \mathbf{y}_k)^2 \tag{14a}$$

$$= (\mathbf{x}_2' \mathbf{y}_k)^2 \tag{14c}$$

$$= (\mathbf{x}_2' \mathbf{H}_2 \mathbf{c}_k)^2 \tag{14d}$$

$$= (\mathbf{d}' \mathbf{c}_k)^2 \tag{15}$$

Comparing Eq. 15 and Eq. 6, we see that the numerator of both Eq. 6 and Eq. 13a are identical.

Now we move to the denominator of Eq. 13a. As in the numerator discussion above, we would now like to put the denominator in a form that is similar to that of the denominator of Eq. 6. Without loss of generality, we can assume that the filtered ACB excitations, $\mathbf{y}_\tau^i = \mathbf{H}_1\mathbf{c}_\tau^i$, are orthogonal. If they are not orthogonal, these can be easily orthogonalized by Gram-Schmidt orthogonalization process. Since the orthogonalization does not change the linear span of $\mathbf{y}_\tau^i$, there is no effect on the overall ACB contribution $\mathbf{y}_\tau$.

Note that the above assumption makes $\mathbf{R}_\tau$ a diagonal matrix. Let R_τ^i be the diagonal elements. Now the denominator of Eq. 13a becomes

$$\mathbf{c}_k' \mathbf{\Phi}\mathbf{c}_k - \sum_{i=1}^{n} \frac{(\mathbf{y}_k'\mathbf{y}_\tau^i)^2}{R_\tau^i} \tag{16}$$

$$\mathbf{c}_k' \mathbf{\Phi}\mathbf{c}_k - \sum_{i=1}^{n} \frac{(\mathbf{c}_k'\mathbf{H}_2'\mathbf{y}_\tau^i)^2}{R_\tau^i} \tag{16a}$$

Now define $\mathbf{z}_i = \mathbf{H}_2'\mathbf{y}_\tau^i$ as the vectors obtained from backward filtering of the filtered ACB excitation vectors. Thus the denominator in Eq. 13 can be written as:

$$\mathbf{c}_k' \left(\mathbf{\Phi} - \sum_{i=1}^{n} \frac{\mathbf{z}_i\mathbf{z}_i'}{R_\tau^i}\right)\mathbf{c}_k \tag{17}$$

The denominator in Eq. 13 can also be written in the form $\mathbf{c}_k' \tilde{\mathbf{\Phi}}\mathbf{c}_k$, where $\tilde{\mathbf{\Phi}} = \mathbf{\Phi} - \sum_{i=1}^{n} \frac{\mathbf{z}_i\mathbf{z}_i'}{R_\tau^i}$. Hence

$$k^* = \underset{\mathrm{k}}{\operatorname{argmax}} \left\{ \frac{(\mathbf{d}'\mathbf{c}_k)^2}{\mathbf{c}_k'\tilde{\mathbf{\Phi}}\mathbf{c}_k} \right\} \tag{18}$$

Since the form of Eqs 6 and 18 are generally the same, the terms $\mathbf{d}$ and $\tilde{\mathbf{\Phi}}$ can be pre-computed, and any existing sequential optimization algorithm may be transformed to a joint optimization without modification to the search algorithm

Going back to Eq. 18, if the vector $\mathbf{Y}_\tau = \mathbf{0}$, 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.

An optimization expression in Eq. 6 and Eq. 18 requires many comparisons of type $p/q < r/s$. In a typical digital signal processor, a multiply has a much lower complexity than a divide. Hence, an equivalent comparison of the form $p{\cdot}s<r{\cdot}q$ is preferred. This requires that the denominators q and s should have the same sign (either positive or negative). It can be verified that the numerator and denominator in Eq. 6 are non-negative.

Since a is a diagonal element of the inverse of a symmetric positive definite matrix $\mathbf{R}$ (inverse is also positive definite), it cannot be negative. Thus, the denominator ($1/a$) is Eq. 18 is also non-negative. Therefore, the equivalent comparisons $p{\cdot}s<r{\cdot}q$ can be performed.

2.1. Complexity

Since the numerator in both Eq. 6 and Eq. 18 are identical, the complexity increase is only from the calculation of the denominator. Note that the matrix modification in the denominator requires an n L-dimension vectors Gram-Schmidt orthogonalization, n backward convolutions, and translation of the correlation matrix $\mathbf{\Phi}$ by n rank 1 matrices $\mathbf{z}_i\mathbf{z}_i'/R_\tau^i$. Since $L = 64$ is significantly larger than the multi-tap filter order n (≤ 3), we can neglect the complexity of Gram-Schmidt orthogonalization and other operations which are linear in L and limit our complexity calculations to operations which are quadratic in L. The backward filtering operation and matrix translations are both quadratic in L. The complexity of backward convolutions is $nL(L+1)/2$. Since the matrices are symmetric, complexity of matrix translation is also $nL(L+1)/2$. Thus the overall complexity increase is approximately $nL(L+1)$. For a narrow band speech coder with $L = 40$, $n = 2$ which means around 3280 extra operations per 5 ms subframe. This is around 14% of the typical FCB search complexity, which is presumed to be around 5M operations/sec. For wideband coder with $L = 64$, $n = 1$, 4160 extra operations per 5 ms subframe are needed, which is around 10% of a typical wideband coder's FCB search complexity (presuming 8M operations/sec). Thus, the complexity is considerably less than the method proposed in [1].

2.2. Discussion

The optimizations process in Eq. 18 assumes that no constraints are placed on ACB gains (multi-tap filter weights). In practical speech coders, the ACB gains are constrained such as bounding the ACB gain between 0.0 and 1.2 and limiting the multi-tap filter to have low pass filter characteristics [4,6]. The sequential search handles this by placing the constraints during the ACB search and decoupling the FCB search from the ACB parameters. We use an ad-hoc approach to handle these constraints. We define constraint handling parameters $0 \leq \alpha_i \leq 1$, and modify

$$\tilde{\mathbf{\Phi}} = \mathbf{\Phi} - \sum_{i=1}^{n} \alpha_i \frac{\mathbf{z}_i\mathbf{z}_i'}{R_\tau^i} \tag{19}$$

When the ACB gains obtained during the ACB search were close to the constraints or outside the constraints, these parameters are made significantly less than unity.

4. RESULTS

For comparisons, the proposed joint optimization was included in the FCB search algorithms of adaptive multi-rate wideband (AMR-WB) coder [5]. Since the purpose of codebook search is to minimize the weighted synthesis error, weighted signal to noise ratio (WSNR) is used for comparisons. The WSNR is defined as the ratio of the energy of the weighted speech to the energy of weighted synthesis error. We use a five minutes wideband speech having sentence pairs spoken alternatively by male and female talkers.

First we consider the case where n = 1. The ACB is constrained to be between [0.0, 1.2]. To handle this, if the initial ACB gain is less than 0.0, then the constraint handling parameter α was chosen to be 0.16. If the gain does not lie in the interval [0.1, 1.15] then α was chosen to be 0.302 for 12.65K mode and 0.25 for 8.85K mode. Otherwise α was set to unity. The results are shown in Table 1. Note that there is a gain of around 0.08 dB between sequential optimization and joint optimization.

Now we consider the case where n = 2. For this a multi-tap filter with a sub-sample resolution delay parameter [4] is integrated in the AMR-WB coder. A *symmetric* multi-tap filter of order 3 was used for simulation purposes, hence n = 2. The taps of the multi-tap filter are also constrained to have low pass characteristics.

If all α_i (Eq. 19) are one then the any set of orthogonal vectors will produce same result. Since the multi-tap filter is constrained and the overall ACB gain is also constrained, we restrict α_i to be less than one. We choose $\mathbf{y}_\tau$ (Eq. 2) as one of the vector and the other is the vector orthogonal to $\mathbf{y}_\tau$ in the span of vectors $[\mathbf{y}_\tau^1, \mathbf{y}_\tau^2]$. We considered two scenarios

- J1. The parameter α_2 is set to zero thus using a joint optimization equivalent to that of n = 1.
- J2. The parameter α_2 = 0.902 when all the constraint were inactive otherwise α_2 = 0.36 for the 8.85K coder 0.205 for the 12.65 K coder.

For both J1 and J2, the parameter α_1 is chosen as in the case of *n=1*. The filter weights and FCB gain were not quantized. The results are shown in Table 2. The gain between sequential optimization and J2 is around 0.1 dB and the gain from J1 to J2 is only 0.03 dB.

5. CONCLUSIONS

A low complexity method for joint optimization of ACB multi-tap filter taps along with FCB excitation and FCB gain has been proposed. This method has complexity advantages over prior methods, and incurs only a 10% to 15% increase in complexity over similar sequential methods with a gain of around 0.1 dB which is equivalent to about 200 bits/s. Furthermore, this method can be easily incorporated into existing ACELP type speech coders through simple translations of the correlation matrix. The search algorithm remains unchanged. This technique is part of the 3GPP2 VMR-WB Rate Set 1 Speech Codec Standard [6].

Mode	Gain Bits	WSNR(dB) Sequential	WSNR(dB) Joint
8.85 K	6 bits	9.04	9.11
8.85 K	Unquantized	9.13	9.21
12.65 K	7bits	11.21	11.29
12.65 K	Unquantized	11.37	11.48

Table 1: Sequential vs. joint optimization in 8.85K and 12.65K mode of AMR-WB coder when the multi-tap filter order is one.

Mode	WSNR(dB) Sequential	WSNR(dB) J2	WSNR(dB) J2
8.85 K	9.59	9.66	9.69
12.65 K	11.93	12.01	12.04

Table 2: Sequential vs. joint optimization in 8.85K and 12.65K mode of AMR-WB coder when multi-tap filter order is two.

6. REFERENCES

[1] Woodard, J. P. and Hanzo, L., "Improvement of analysis by synthesis loop in CELP codecs," *IEE 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., and Massaloux, D., "A toll quality 8 Kb/s speech codec for personal communications system," *IEEE Tran. on Vehicular Technology,* 808-816, Aug. 1994.

[4] Jasiuk M. A., Ramabadran T., Mittal U., Ashley J. P., and McLaughlin M. J., "A technique of multi-tap LTP filter using sub-sample resolution delay," *accepted for presentation at ICASSP 2005.*

[5] "AMR Wideband Speech Codec Technical Specifications," Document 3GPP TS 26.190, Version 5.1.0, Dec. 2001.

[6] "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Option 62 and 63 for Spread Spectrum Systems," Document 3GPP2 C.P0052-A, Version 0.3, December 10, 2004.