S4.4

IMPROVED SPEECH QUALITY AND EFFICIENT VECTOR QUANTIZATION IN SELP

W. B. Kleijn, D. J. Krasinski, and R. H. Ketchum

AT&T Bell Laboratories
Naperville, IL

Abstract

A SELP (Stochastically Excited Linear Prediction) algorithm consisting of a two-stage vector quantization using an adaptive codebook and a stochastic codebook is described. The adaptive codebook quantization is similar to a closed loop long-term prediction filter if the predictor delay is more than one frame length. The performance of the adaptive codebook procedure is improved by extending its codebook to include candidate vectors constructed from past synthetic excitation displaying a high level of periodicity. A new algorithm is introduced, which through increased symmetry of the error criterion significantly reduces the computational effort required for the search through the adaptive codebook. The algorithm can also be used for stochastic codebooks consisting of overlapping candidate vectors. It is shown that for stochastic codebooks in which neighboring candidates overlap for all but two samples, the quantization performance is as high as for codebooks containing fully independent candidate vectors.

1. Introduction

The algorithm described here falls into a general class of speech coding algorithms [1] which use analysis-by-synthesis procedures to find an appropriate excitation function for a synthesis filter determined from a linear predictive (LP) analysis procedure. The first algorithm of this class to be introduced was the Multi-Pulse Linear Prediction (MPLP) algorithm [2].

The MPLP algorithm can be viewed as a multiple-stage vector quantization of the excitation sequence with a deterministic codebook consisting of candidate vectors each containing a single pulse. The MPLP algorithm was later enhanced with an additional stage, which consists of a search for the best excitation vector through the past synthetic excitation [3]. This search through a continuously updated recent past can be interpreted as a search through an adaptive codebook. The adaptive codebook greatly improves the synthetic speech quality produced by the MPLP algorithm, removing much of its gravelly character. Subsequently, it was shown that the coder performs well even if the multi-pulse stages are removed and only the adaptive codebook is used [4].

Instead of an adaptive codebook or a deterministic codebook, a codebook containing random sequences can also be employed [5]. We designate coders which use a stochastic codebook as Stochastically Excited Linear Prediction (SELP). If this stochastic codebook is used for single-stage vector quantization in combination with a short-term LP filter, good speech quality at low bit rates can only be obtained for extremely large codebooks which require very high computational efforts. Otherwise, the stochastic codebook is found to result in a comparatively noisy synthetic speech. A long-term predictor or an adaptive codebook can be used to minimize this problem.

A long-term predictor describes the periodicity of the speech signal efficiently, and has been applied in open-loop and closed-loop form. In the former case the long-term predictor is determined by analyzing the original speech signal only, in the latter case the optimal long-term predictor is determined with an analysis-by-synthesis procedure. From this description, it can be noted that the closed-loop long-term predictor and the adaptive codebook are very similar. As will be shown in section 2 they are not identical; in particular, the fast algorithm described in section 3 can be applied to the search through the adaptive codebook but not for the determination of a long-term predictor. For this reason the algorithm described here combines an adaptive codebook with a stochastic codebook.

In the past the adaptive codebook procedure has either been used in combination with other codebooks, or with relatively short frame lengths [4]. This reflects the fact that the performance of the adaptive codebook procedure deteriorates when the frame length is more than the length of a pitch period. In section 2, the adaptive codebook will be extended to obtain higher performance for longer frame lengths. The improvement provided by this extension is particularly noticeable at the onset of voiced segments of speech.

The computational effort for finding the best entry from the stochastic codebook is large. Previously, much work has focused on the reduction of the computional effort of the stochastic codebook search. Many different approaches have been described, such as the use of sparse codebooks [6], [7] and evaluation of the error criterion in a transform domain [8]. The adaptive codebook procedure benefits if the considered synthetic history (from which the present frame is reconstructed) is increased in length; however, this means that the computational effort grows significantly. A new fast method which applies both to the stochastic and the adaptive codebook will be introduced in section 3. The procedure uses a modified error criterion, which displays a high level of symmetry. This error criterion has been found, with both objective (signal-to-noise ratios) and subjective (listening tests) measures, to work as well as or better than the conventional error criterion. The new fast algorithm takes advantage of the symmetry with a minimum of overhead, computing the search through the adaptive codebook faster than previously possible.

2. Improved Quality of the SELP Algorithm

2.1 The Basic Algorithm. Consider a frame of speech signal described by the vector **s**. Regardless of the past synthetic excitation we want to construct an excitation sequence **t** such that filtering **t** through the LP filter results in the speech signal **s** for the present frame. The convolution of the excitation sequence **t** with the impulse response of the filter can be written as a matrix multiplication **Ht**, where **H** is a Toeplitz matrix containing the impulse response of the filter in its first column:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & \vdots \\ \vdots & h_1 & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ h_{N-1} & h_{N-2} & \cdots & h_0 \end{bmatrix}. \qquad (1)$$

If $\mathbf{z}$ is the zero-input response of the system for this frame, the speech signal $\mathbf{s}$ can written as:

$$\mathbf{s} = \mathbf{Ht} + \mathbf{z}. \tag{2}$$

The SELP algorithm implements a vector quantization procedure of the target excitation vector $\mathbf{t}$. However, it must be kept in mind that the goal is to accurately model the speech signal $\mathbf{s}$, and not the target excitation $\mathbf{t}$. An obvious method to accomplish this is by replacing $\mathbf{t}$ in equation (2) with the set of optimally scaled candidate vectors $\mathbf{r}_i$ contained in the codebook and determining which candidate vector results in the best synthetic speech signal. The matrix $\mathbf{H}$ provides a weighting of the $\mathbf{r}_i$ and $\mathbf{t}$ vectors in the evaluation of the least squares error criterion:

$$\epsilon = (\mathbf{t}-\mathbf{r}_i)^T\mathbf{H}^T\mathbf{H}(\mathbf{t}-\mathbf{r}_i), \tag{3}$$

The described spectral weighting of the vector quantization of the target vector $\mathbf{t}$ results in an equal expectation value for the error over the entire spectrum of the synthetic speech signal. It is known, however, that the quantizer noise is masked by the formants of the speech signal, suggesting that more of the quantizer noise should be put under the formants [9]. To diminish the perceived noise signal resulting from the vector quantization procedure, the relative weighting of the formant regions can be increased slightly through deemphasizing the spectrum contained by the $\mathbf{H}$ matrix. This can be accomplished by moving the poles of the spectrum inward by a constant factor (usually chosen to be around 0.8).

When a multiple-stage vector quantization procedure is used, the error criterion described by (3) is applied at each stage. After the selection of the optimally scaled winning candidate vector $\mathbf{r}_j$ for a particular quantization stage, the target for the next vector quantization stage is obtained by subtracting $\mathbf{r}_j$ from the current target vector.

The $\mathbf{H}$ matrix of (1) results in an error criterion which displays only a limited amount of symmetry. The development of fast algorithms benefits from increased symmetry. Another disadvantage of the described error criterion is that the last sample of the candidate vector will affect only the last synthetic speech sample of the present frame, while the first sample of the candidate vector will affect all synthetic speech samples in the present frame. The uneven weighting of the samples in the candidate vector will result in accurate matching of the target at the beginning of the frame, and more erratic behavior for the last samples of the target vector. This erratic behavior is then compensated for with the accurate matching of the early samples of the next target vector.

A different error criterion results in more evenly weighting of the target vector samples. We maintain essentially the same spectral weighting, but truncate the impulse response after $\mathbf{R}$ samples, using the entire response of length $\mathbf{R}$ for each sample of $\mathbf{r}_i$ and $\mathbf{t}$. The matrix $\mathbf{H}$ becomes:

$$\mathbf{H} = \begin{bmatrix} h_0 & 0 & \cdot & \cdot & 0 \\ h_1 & h_0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ h_{R-1} & h_{R-2} & \cdot & \cdot & \cdot \\ 0 & h_{R-1} & \cdot & \cdot & h_0 \\ \cdot & \cdot & \cdot & \cdot & h_1 \\ \cdot & \cdot & \cdot & h_{R-1} & h_{R-2} \\ 0 & \cdot & \cdot & 0 & h_{R-1} \end{bmatrix}. \tag{4}$$

The error criterion of equation (3) can be employed with this modified $\mathbf{H}$ matrix. The matrix $\mathbf{H}^T\mathbf{H}$ is still NxN but the vectors $\mathbf{H}(\mathbf{r}_i-\mathbf{s})$ are now of length N+R. Note that procedures described in [10] and [1] also employ an algorithm with a non-square $\mathbf{H}$ matrix, but for the special case where R=N.

A major advantage of the modification of the error criterion is the symmetry created in the $\mathbf{H}^T\mathbf{H}$ matrix, which is now a symmetric Toeplitz band matrix. If expression (1) is used in the error criterion then the $\mathbf{H}^T\mathbf{H}$ matrix is symmetric only. The added symmetry of the modified error criterion facilitates the fast algorithm described in the section 3.

2.2 An Improved Excitation Function. If the long-term predictor is used, then the spectral weighting matrix $\mathbf{H}^T\mathbf{H}$ must be modified to include the effect of this additional filter. This modification does not affect the weighting matrix if the long-term prediction delay is longer than the frame length employed for the stochastic codebook. In this case, the closed-loop long-term prediction is identical to the adaptive codebook procedure. For longer frame lengths, which are more efficient from a distortion-theoretic viewpoint [11], these two procedures are not identical. The adaptive codebook procedure allows a more straightforward implementation since the $\mathbf{H}^T\mathbf{H}$ matrix does not contain effects from the long-term predictor. The main advantage is, however, the fast algorithm described in section 3 which can only be used for the search through the adaptive codebook and not for the determination of the optimal closed pitch loop.

A disadvantage of the adaptive codebook procedure appears to be that the most recent candidate vector for the present frame is at least one framelength prior to the present frame. At the onset of a voiced section of speech the most recent candidate vector might contain only a single pitch pulse, while two pitch pulses might be required. Thus, periodicity at shorter than a single frame length is not well modeled. Short frame lengths are not desirable for transmission efficiency, however, since the excitation information can be encoded more efficiently if longer frame lengths are used. The following procedure alleviates this problem.

Consider a candidate vector which starts $\mathbf{p}$ samples prior to the present frame, where $\mathbf{p}$ is less than $\mathbf{N}$, the frame length. Thus, only the first $\mathbf{p}$ samples of this candidate vector are defined. If this candidate vector (which we imagine to have been completed in some as yet undetermined fashion) is the winning vector, then these first $\mathbf{p}$ samples would occupy the first $\mathbf{p}$ locations of the present frame. Neglecting scaling, these samples repeat the previous set of samples with a periodicity equal to $\mathbf{p}$. An obvious extension of the candidate vector is to repeat its first part with the same periodicity $\mathbf{p}$. In other words, the previously undefined part of the candidate vector is constructed by repeating its defined part (of length $\mathbf{p}$) with a periodicity of $\mathbf{p}$. The creation of these "virtual" candidate vectors enhances the build-up of periodicity at the start of a voiced region as illustrated in Fig. 1, and the modeling of voices with a high pitch. The extended adaptive codebook differs from the closed-loop long-term predictor in the relative scaling of the repeated waveform.

3. The Fast Algorithm

3.1 Description of the Algorithm. The original SELP algorithms required a very high computational effort for the determination of the excitation vector. For a fast computation of the error criterion (3) it is useful to write it in the following form:

$$\epsilon = \mathbf{t}^T\mathbf{H}^T\mathbf{Ht}-2\mathbf{r}_i^T\mathbf{H}^T\mathbf{Ht}+\mathbf{r}_i^T\mathbf{H}^T\mathbf{Hr}_i. \tag{5}$$

When searching through the codebook the first term of equation (5) does not vary, and can be ignored. For the second term the vector $\mathbf{H}^T\mathbf{Ht}$ can be evaluated once, resulting in N remaining instructions per candidate vector.

The third term of equation (5) requires more effort. Since the neighboring candidates of the adaptive codebook overlap for all but a single sample, the response of the synthesis filter to a particular candidate can be computed from a neighboring candidate's response by adding and subtracting the response of the two samples not shared by these neigboring candidates [4]. This procedure applies for both the modified error criterion and the original error criterion. The computational effort per candidate in this codebook is $\mathbf{2*R}$ instructions to compute the response to the non-shared samples, and N or N+R instructions for evaluating the sum of the

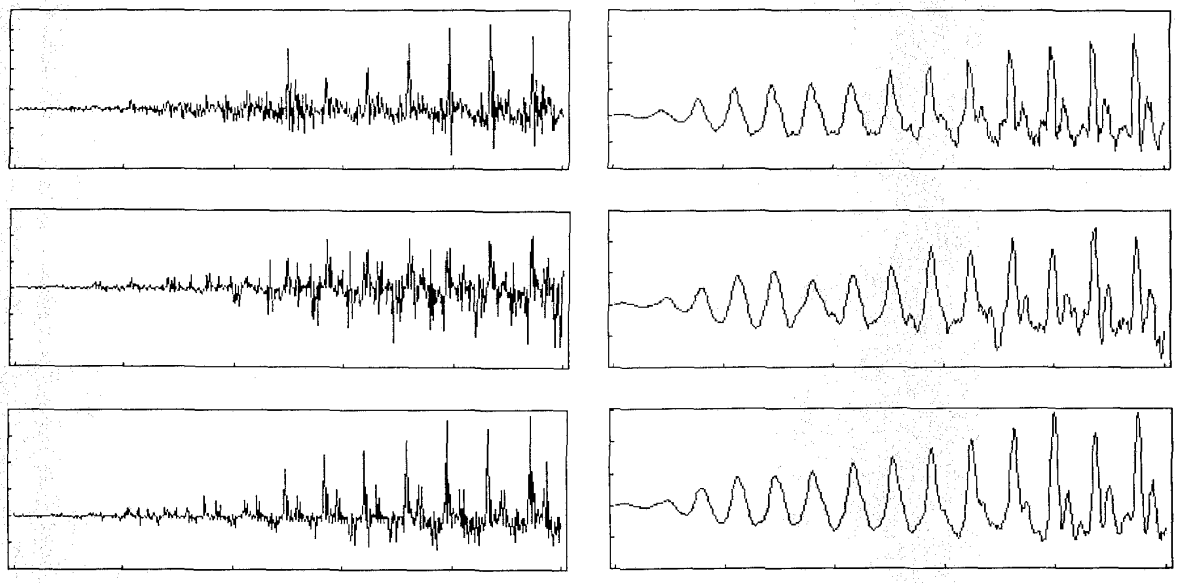

Figure 1

An illustration of the effect of the extended adaptive codebook on a 62.5 ms section of speech.
Top: excitation yielding original speech; original speech
Middle: synthetic excitation without extended adaptive codebook; associated synthetic speech
Bottom: synthetic excitation with extended adaptive codebook; associated synthetic speech

squares, depending on the error criterion used. At a savings of **R** instructions, the responses of all samples can be stored, but the overhead to organize the data storage may make direct computation more desirable.

The algorithm described in this section improves significantly on the above described fast procedure for the third term. Instead of evaluating the responses to each candidate sequence, the new procedure takes advantage of the structure of the $\mathbf{H^TH}$ matrix. It is advantageous to define a shifting matrix **S** of dimension NxN:

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & . & . & 0 \\ 0 & 0 & 1 & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & 0 & 1 \\ 0 & . & . & . & 0 & 0 \end{bmatrix}. \qquad (6)$$

In addition, we define a masking matrix $\mathbf{I_k}$ of dimension NxN:

$$\mathbf{I_k} = \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & 1 & 0 & . \\ . & . & . & 0 & 0 & . \\ . & . & . & . & . & . \\ 0 & . & . & . & . & 0 \end{bmatrix}, \qquad (7)$$

where the last nonzero element appears in the kth row. A candidate vector $\mathbf{r_{i+1}}$ can now be expressed in terms of the prior candidate $\mathbf{r_i}$ as follows:

$$\mathbf{r_{i+1} = Sr_i + (I-I_{N-1})r_{i+1}}. \qquad (8)$$

The object is to obtain a recursion which relates the energy $\mathbf{E_{i+1}=r_{i+1}^TH^THr_{i+1}}$ to the energy of the preceding candidate $\mathbf{E_i=r_i^TH^THr_i}$. Realizing that $\mathbf{S^TAS=(I-I_1)A(I-I_1)}$, we obtain after some algebra:

$$\mathbf{E_{i+1} = E_i + 2\left[r_{i+1}^T(I-I_{N-1})H^THSr_i - r_i^T(I-I_1)H^THI_1r_i\right] -}$$

$$\mathbf{r_i^TI_1H^THI_1r_i + r_{i+1}^T(I-I_{N-1})H^TH(I-I_{N-1})r_{i+1}}. \qquad (9)$$

The computational effort required to evaluate this expression can be evaluated in the following manner. The vector $\mathbf{r_{i+1}^T(I-I_{N-1})}$ contains only one nonzero entry, thus the term $\mathbf{r_{i+1}^T(I-I_{N-1})H^THSr_i}$ requires a total of **R** instructions. Similarly the term $\mathbf{r_i^T(I-I_1)H^THI_1r_i}$ requires R instructions since $\mathbf{I_1r_i}$ also contains **R–1** zero entries. The multiplication by 2 requires one more instruction. The terms $\mathbf{r_i^TI_1H^THI_1r_i}$ and $\mathbf{r_{i+1}^T(I-I_{N-1})H^TH(I-I_{N-1})r_{i+1}}$ each require 2 instructions. Thus, the recursive evaluation of (9) requires **2R+3** instructions. This compares to **N+2R** or **N+3R** with the straightforward recursive algorithm described earlier. For example, in a 4.8 kb/s system with a frame length **N=60** and an impulse response length of **R=25**, equation (9) requires 53 instructions per recursion, while the conventional algorithm requires *110* or *135 instructions per recursion*. Note that if **2R+3<N**, then the second term of equation (5) requires more computational effort than its third term.

A similar procedure can be applied to the virtual search as well, reducing its computational effort significantly.

3.2 Application of the Fast Algorithm to the Stochastic Codebook. The above recursive procedure can also be employed for a stochastic codebook if it is of an overlapping nature. Overlapping stochastic codebooks, which have been employed previously [7] are also of interest for their minimal storage requirements. It appears that the dependence of neighboring candidate vectors may deteriorate the performance of the stochastic codebook. This dependence diminishes with increasing shift between the candidates contained in the codebook. Table I provides segmental signal to noise ratios over a set of 19 sentences for a representative

SELP algorithm run with a variety of shifts at two codebook sizes. It is seen that for larger codebooks a shift of 2 results in performance identical to a fully independent codebook. Thus, the application of the fast algorithm does not adversely affect performance.

The efficiency of the algorithm is further enhanced, if a sparse stochastic codebook is used. The recursion of (9) is simplified whenever either of the non-shared elements of the neighboring candidates is zero. If both non-shared elements are zero (9) reduces to $\mathbf{E}_{i+1}=\mathbf{E}_i$. We have found that a sparse codebook (approximately 90% zero samples) results in marginally increased subjective performance.

Table I: Overlapping Stochastic Codebook		
Segmental Signal to Noise Ratios		
shifts	32 candidates	128 candidates
1	10.4	11.9
2	10.8	12.1
4	11.2	12.1
32	11.2	12.1
128	-	12.1

4. An Implementation of the SELP Algorithm

The described algorithm has been implemented at both 4.8 and 8.0 kb/s rates, using an 8 kHz sampling rate. To provide high speech quality, efficient encoding of the Linear Predictive Coding (LPC) parameters is essential. However if differential coding of LPC parameters is used, the coder will be highly sensitive to bit errors in the channel. As a compromise we employ an interpolation scheme. When using interpolative (and predictive) procedures for the LPC parameters, it is essential to employ a description which behaves smoothly in time, such as line spectral frequencies. In the interpolation procedure, a conventional scalar quantization is performed for every second determination of the LPC parameters (at 4.8 kb/s the LPC parameters are determined every 22.5 ms). The quantized values of these LPC parameters are interpolated to the intermediate LPC frames (note that these frames are not identical to previously mentioned excitation quantization frames). The difference vector between the set of interpolated LPC parameters, and the actual LPC parameters is then vector quantized. In sections of speech where the spectrum is well-defined, i.e. where the formants do not vary rapidly, this procedure will result in high accuracy for the intermediate LPC frames, where vector quantization is applied. If the formant structure is strongly time-dependent, and the spectrum is ill-defined, the procedure will result in less accurate results, but under such circumstances a high degree of accuracy is not required.

The LPC parameters, which are updated less often than the excitation information, are then interpolated, resulting in a new **H** matrix every **N** samples. Our implementation of the LPC quantization requires 1.2 kb/s for the 4.8 kb/s coder and 2.0 kb/s for the 8.0 kb/s coder.

The quantization of the optimal scaling factor for each candidate vector of the stochastic codebook is performed before the evaluation of the error criterion. The scaling factor of the chosen stochastic codebook entry is highly correlated with the energy of the excitation sequence obtained from the adaptive codebook. The appropriate renormalization of the stochastic codebook improves the quantizer performance.

A postfilter is employed to remove some of the perceived noise of the coding algorithm. A modified version of the postfilter described by [12] is used. The goal of postfiltering is to emphasize the formants of the synthetic speech signal. The effect of the original all-pole postfilter [13] was to not only emphasize the formants of the filter but also the overall spectral tilt of the speech signal. The pole-zero postfilter removes this problem to a large extent. The poles and zeros of this postfilter are along the same radials as the poles of the infinite impulse response filter for the synthesizer. Both the poles and the zeros are moved inward from the synthesizer poles. The function of the zeros (which are further inward) is to cancel the effect of the postfilter poles far from their actual location. Thus, the effect of the pole-zero postfilter is highly localized. The spectrum of the pole-zero postfilter will contain only a minor overall tilt. In [12], an additional filter with a single zero at 0.5 is applied to remove the remaining tilt. Since the tilt of the spectrum varies, we have found it beneficial to adapt this filter and locate the zero at $\mathbf{0.5k_1}$, where $\mathbf{k_1}$ is the first reflection coefficient computed from the quantized LPC parameters. We have found that an all-pole filter with its single pole at $\mathbf{-0.5k_1}$ is subjectively of slightly higher quality, although the objective performance deteriorates.

REFERENCES

[1] Atal, B.S., "High-Quality Speech at Low Bit Rates: Multi-Pulse and Stochastically Excited Linear Predictive Coders" *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Tokyo, 1681-1684, 1986.

[2] Atal, B.S. and J.R. Remde, "A New Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Paris, 614-617, 1982.

[3] Singhal S. and B.S. Atal, "Improving Performance of Multi-Pulse LPC Coders at Low Bit Rates", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, San Diego, 1.3.1-1.3.4, 1984.

[4] Rose, R.C. and T.P. Barnwell, "Quality Comparison of Low Complexity 4800bps Self Excited and Code Excited Vocoders", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Dallas, 1637-1640, 1987.

[5] Atal B.S. and M.R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rates", *Proc. of ICC*, Amsterdam, 1610-1613, 1984.

[6] Davidson, G. and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding", *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Tokyo, 3055-3058, 1986.

[7] Lin, D., "New approaches to Stochastic Coding of Speech Sources at Very Low Bit Rates", in *Signal Processing III: Theories and Applications*, I.T. Young et al. eds., Elsevier, 445-447, 1986.

[8] Transcoso, I.M. and B.S. Atal, "Efficient Procedures for Finding the Optimum Innovation in Stochastic Coders" *Proc. Int. Conf. Acoust., Speech and Sign. Process.*, Tokyo, 2379-2382, 1986.

[9] Atal, B.S. and M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria" *IEEE Trans. Speech Signal Proc. Vol. ASSP-27, no 3*, 247-254, 1979.

[10] Kroon, P., Ed F. Deprettere, and R.J. Sluyter, "Regular-Pulse Excitation - A Novel Appoach to Effective and Efficient Multipulse Coding of Speech", *IEEE Trans. Acoust. Speech Signal Proc.*, Vol. ASSP-34, No 5, 1986.

[11] Berger, T., *Rate Distortion Theory*, Prentice-Hall, New Jersey, 1971.

[12] Chen, J.H. and Gersho, A., "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering", *Proc. Int. Conf. Acoust., Speech and Sign. Process. Dallas*, 2185-2188, 1987.

[13] Ramamoorthy, V. and Jayant, N.S., "Enhancement of ADPCM Speech by Adaptive Postfiltering", *AT&T Bell Labs. Tech. J.*, 1465-1475, 1984.