

AN ADAPTIVE MULTI-RATE SPEECH CODER FOR DIGITAL CELLULAR SYSTEMS*

Ahmed J. Jameel, You Xiaohu, Wang Ling and Gao Xiqi
(National Mobile Communications Research Laboratory)

(Department of Radio Engineering, Southeast University, Nanjing 210096, China)

Abstract

Low cost speech coding is very important for mobile communications. In this paper we propose a low-complexity but efficient speech coder. The coder is an adaptive multi-rate ACELP coder based on the RCELP paradigm operating at bit-rates of 7.6, 6.15, 4.85 and 0.8 kbits/s. The coder provides seamless switching between rates without annoying artifacts on 20 ms frame boundaries. The listening tests show that the subjective quality is better than that of EVRC. The performance of the encoder under AWGN and Rayleigh channels show that it is robust against channel errors.

Keywords: AMR, Speech coding, CELP and EVRC.

1. Introduction

The rapid increase in the use of mobile communications has recently caused a shortage of available radio frequencies. Lower bit-rate coding of speech by digital signal processing is one of the most promising ways to make more channels available. Speech coders for digital mobile communications must keep good quality in various severe conditions such as different speakers, channel errors, various input speech levels, and background noise [1].

With the deployment of digital cellular systems such as 13 kb/s GSM RPE-LTP, 8 kb/s fixed rate IS54 VSELP and 8 kb/s variable rate IS96 QCELP, it is becoming clear that these coders are not as robust as originally expected. As a result there has been renewed interest in coders that provide toll quality performance with regard to various input signals and multi-coding rate [2].

Linear prediction analysis-by-synthesis (LPAS) coders using a codebook approach are commonly referred to as code-excited linear prediction (CELP) coders [3]. Although many improvements were made in the coding efficiency of LPAS coders, the basic analysis-by-synthesis paradigm has not changed since the introduction of CELP. However, at low bit rates, the matching of an original waveform becomes a severe constraint in improving the coding efficiency further. In generalized linear prediction analysis-by-synthesis (GLPAS) coding, the original speech signal is modified such that; it can be coded more effectively by the analysis-by-synthesis coder. The modification of the original can be done either by time warping [8] or time shifting [6]. Time warping guarantees a continuous evolution of the waveforms but is computationally expensive. Time shifting is of a much lower complexity, but it introduces time-shift boundaries. It is important to ensure that time-shifting boundaries are avoided in excitation signal segments where the power is large (i.e. pitch pulses). If the discontinuities are made to fall consistently in low-energy regions, the performance of the time-shifting procedure is the same as that of the time-warping procedure.

This paper is arranged as follows: First we give a general description of the algorithm. In section 3, adaptive codebook search procedure is described. In section 4, fixed codebook – search and structure are presented. In section 5, the performance of the system is evaluated and we conclude in section 6.

2. General Description of the Algorithm

This coder is a multi-rate ACELP coder based on the relaxation CELP (RCELP) paradigm [6]. Unlike conventional CELP codec, RCELP attempts to match a modified speech residual signal generated by a time-warped version of the original residual that conforms to a simplified pitch contour. As a result, the pitch information is transmitted over a frame instead of a subframe. Consequently, more bits are allocated

* This work has been supported by the National Natural Science Foundation of China (No. 69725001).

to the fixed codebook encoding and to the channel coding. The coder operates on speech frames of 20 ms corresponding to 160 samples at a sampling rate of 8000 samples per second. After processing the input samples through a second order highpass filter with a cut-off frequency of 140 Hz, tenth order LPC analysis is performed, and the LPC parameters are interpolated and quantized in the line spectral pair (LSP) domain. Each frame of the input speech is divided into four subframes. The use of subframes allows better tracking of the pitch and gain parameters and reduces the complexity of the codebook searches. Table 1 shows the bit allocations for each packet type. At rates 1, 2 and 3, the encoder will apply the RCELP algorithm to match a time-warped version of the original speech residual. If rate 4 is selected, the encoder will not attempt to characterize any periodicity in the speech residual, but instead just characterize its energy contour.

The input speech vector is divided into six segments, and a different set of interpolated LSPs is computed for each corresponding segment. The interpolated LSPs are converted to LPCs. For each segment k , the unquantized, interpolated LSP vector is:

$$\hat{\Omega}(m, k) = (1 - \mathbf{m}_k) \Omega(m-1) + \mathbf{m}_k \Omega(m); \quad 0 \leq k \leq 5, \quad (1)$$

where the interpolator constants, $\{\mathbf{m}\}$, and their corresponding sets of sample indices for each segment of input speech are given in Table 2. The short-term prediction residual is generated by passing the input speech signal through the inverse filter using the appropriate LPCs.

3. Adaptive codebook Search

The adaptive codebook parameters are the pitch delay, delta delay and the adaptive codebook gain (ACB). The pitch delay is estimated through open-loop analysis by maximizing the autocorrelation function of the short-term prediction residual signal using a 20 ms window, and is quantized to 7 bits.

Since the adaptive codebook delay directly affects the periodicity of the speech, bit errors in this parameter affect the perceptual speech quality significantly (much more than the fixed-codebook index). Thus, the RCELP coder, which has fewer bits allocated to the adaptive codebook delay (7 bits per frame) and more to the fixed codebook index, displays increased robustness to channel errors [7]. This performance can be further enhanced for the case of frame erasure by constraining the adaptive codebook delay to be less than 2 ms per frame (this does not affect the reconstructed-speech quality). For a 1-sample resolution at an 8 kHz sampling-rate, the change in the adaptive codebook delay, the delta delay, can then be described with 5 bits per frame. Transmission of both the adaptive codebook delay and the delta delay allows the decoder to reconstruct the correct delay contour even when frames are erased, if no more than one frame is erased at a time.

4. Fixed Codebook – Structure and Search

The fixed codebook is based on an algebraic codebook structure, which has advantages in terms of storage, search complexity, and robustness. The codebook structure is based on an interleaved single-pulse permutation (ISPP) design [9]. The algebraic codebook is a deterministic codebook whereby the excitation code vector is derived from the transmitted codebook index (no need for codebook storage). The codebook is searched on a subframe basis for the best index and gain to minimize the mean-squared weighted error between the original and synthesis speech.

4.1 Algebraic Codebook Structure, Rate 1 and Rate 2

The Rate 1 fixed codebook is a 20-bit and Rate 2 is a 14-bit algebraic codebook. In these codebooks, every codebook vector of length 40 contains at most 5 non-zero pulses for Rate 1 and 3 non-zero pulses for Rate 2. All pulses can have the amplitude +1 or -1. The 40 positions in a subframe are divided into five tracks of eight positions. For Rate 1, each track can have one pulse. For Rate 2, the first four tracks can have two pulses, while the last pulse is placed in the fifth track. The sign of each pulse is quantized with 1 bit and its position is quantized with 3 bits. This gives a total of 20 bits for Rate 1 and 14 bits for Rate 2. The codebook vector, c_k , is constructed according to:

$$c_k(j) = \sum_{i=0}^{N_p-1} s_i \mathbf{d}(j - p_i); \quad 0 \leq j \leq 39, \quad (2)$$

where $\mathbf{d}(j - p_i)$ is a unit pulse at the i -th pulse position p_i of the k -th codevector, s_i is the sign of the i -th pulse, N_p is the number of pulses, and k is the range of all possible code vectors.

4.2 Algebraic Codebook Search

The algebraic codebook is searched by minimizing the mean-squared error between the weighted input speech and the weighted synthesis speech. The perceptual domain target signal $x_w(n)$ is used in the closed-loop fixed-codebook search.

Let c_k be the algebraic codebook vector at index k , the algebraic codebook is searched by maximizing the term:

$$T_k = \frac{C_k}{E_k} = \frac{(\mathbf{d}^t \mathbf{c}_k)^2}{\mathbf{c}_k^t \Phi \mathbf{c}_k}, \quad (3)$$

where $\mathbf{d} = \mathbf{H}' \mathbf{x}_w$ is the cross-correlation between the perceptual domain target signal $x_w(n)$ and the impulse response $h_{wq}(n)$, $\Phi = \mathbf{H}' \mathbf{H}$ is the correlation matrix of the impulse response $h_{wq}(n)$, and \mathbf{H} is a lower triangular Toeplitz matrix with diagonal $h_{wq}(0)$ and lower diagonals $h_{wq}(1), \dots, h_{wq}(39)$, i.e.:

$$\Phi = \begin{bmatrix} h_{wq}(0) & 0 & 0 & \cdots & 0 \\ h_{wq}(1) & h_{wq}(0) & 0 & \cdots & 0 \\ h_{wq}(2) & h_{wq}(1) & h_{wq}(0) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{wq}(39) & h_{wq}(38) & h_{wq}(37) & \cdots & h_{wq}(0) \end{bmatrix}, \quad (4)$$

The cross-correlation vector \mathbf{d} and the matrix Φ are computed prior to the codebook search. The elements of the vector \mathbf{d} are computed by:

$$d(n) = \sum_{j=n}^{39} x_w(j) h_{wq}(j-n); \quad 0 \leq n \leq 39, \quad (5)$$

and the (i, j) -th element of the symmetric matrix Φ is computed by:

$$f(i, j) = \sum_{n=\max\{i, j\}}^{39} h_{wq}(n-i) h_{wq}(n-j); \quad (0 \leq j \leq 39) \text{ and } (0 \leq i \leq 39). \quad (6)$$

The algebraic structure of the codebook allows for very fast search procedures since the innovation vector, c_k , contains only few non-zero pulses. The correlation in the numerator of Equation (3) is given by:

$$C_k = \left(\sum_{i=0}^{N_p-1} s_i d(p_i) \right)^2. \quad (7)$$

The energy in the denominator of Equation 3 is given by:

$$E_k = \sum_{i=0}^{N_p-1} f(p_i, p_i) + 2 \sum_{i=0}^{N_p-2} \sum_{j=i+1}^{N_p-1} s_i s_j f(p_i, p_j). \quad (8)$$

4.2.1 Pre-setting of Pulse Signs

In order to simplify the search procedure, the pulse signs are preset (outside the closed loop search) by considering the sign of an appropriate reference signal. In this case, the signal $e(i)$, given by:

$$e(i) = \sqrt{\frac{\sum_{j=0}^{39} d^2(j)}{\sum_{j=0}^{39} x^2(j)}} x(i) + 2d(i); \quad 0 \leq i \leq 39, \quad (9)$$

shall be used, where $x(i)$ is the residual domain target vector. Amplitude pre-setting shall be done by setting the amplitude of a pulse at position i equal to the sign of $e(i)$. Hence, once the sign signal $s_i = \text{sign}\{e(i)\}$ and the signal $d'(i) = d(i)s_i$ are computed, then the matrix Φ shall be modified by including the sign information, that is, $f'(i, j) = s_i s_j f(i, j)$. The correlation in Equation 7 is now given by:

$$C_k = \left(\sum_{i=0}^{N_p-1} d'(p_i) \right)^2, \quad (10)$$

and the energy in Equation 8 is given by:

$$E_k = \sum_{i=0}^{N_p-1} f'(p_i, p_i) + 2 \sum_{i=0}^{N_p-2} \sum_{j=i+1}^{N_p-1} f'(p_i, p_j). \quad (11)$$

4.2.2 Non-Exhaustive Pulse Position Search

Having preset the pulse amplitudes, the optimal pulse positions shall be determined using an efficient non-exhaustive analysis-by-synthesis search technique. In this technique, the term in Equation 3 is tested for a small percentage of position combinations, using an iterative “depth-first” tree search strategy. Once the positions and signs of the excitation pulses are determined, the codebook vector c_k , shall be built as in Equation 2.

4.3 Algebraic Codebook Structure, Rate 3

A 10-bit algebraic codebook is used for Rate 3 packets. The innovation vector contains 3 non-zero pulses. Each pulse has 8 possible positions, which is coded by 3 bits. The pulse positions and corresponding codewords are given in Table 4. All pulses have fixed signs (+1 for T_0 and T_2 and -1 for T_1). An additional bit, however, is used to change the signs of all three pulses simultaneously, i.e., $s = 1$ indicates polarity inversion, $s = 0$ otherwise. Therefore, the total bits per subframe for Rate 3 is $3 \times 3 + 1 = 10$.

The codebook is searched using techniques similar to that for Rate 1. However, in Rate 3 case, all the pulse position combinations are exhaustively searched by maximizing Equation 3.

5. Performance Evaluation

The proposed speech Coder operates on frames of 160 samples (20 ms) and produces 152 bits per frame for rate 1, 123 bits per frame for rate 2, and 97 bits per frame for rate 3. The LPC parameters are determined every 20 ms, and quantized with a weighted split vector LSP quantizer. The excitation is determined every 5 ms and for each subframe the coder provides pitch information, a fixed-codebook index, and gain information.

A sensitivity analysis was used to rank the bits in terms of sensitivity. It was found that the LSP index and the overall pitch information are the most sensitive. The codebook gains are of medium sensitivity and the pulse positions and pulse sign are the least sensitive. By using the objective and subjective test, we can divide the bits of the proposed encoder into three different classes according to their effect on the coder performance. The three classes, are presented in Table 5, where class 1 is the most sensitive, class 2 comprises the medium sensitive bits, that must be protected against channel errors, while class 3 is the robust class and can be used without protection with some acceptable degradation in speech quality.

Many objective error measures exist [4], but the signal-to-noise ratio (SNR) in dB is a commonly used measure and is defined as:

$$SNR = 10 \times \log_{10} \frac{\sum_{i=0}^{N-1} x^2(i)}{\sum_{i=0}^{N-1} [x(i) - y(i)]^2}. \quad (12)$$

where $x(n)$ and $y(n)$ are the input and the output of the coder, respectively. Since speech is a nonstationary signal with many high and low energy sections that are perceptually relevant, a better approach is to compute the SNR for shorter segments and to compute the statistics of this local SNR value. This measure is referred to as the *segmental* SNR and is defined as:

$$SNR_{SEG} = \frac{1}{K} \sum_{k=1}^K SNR_k. \quad (13)$$

The coder was tested under two conditions. The first one is under error free conditions and the second one under channel errors. Table 6, contains the objective test results under error free conditions for the proposed encoder compared with the standard EVRC encoder for three languages using the same speech sentences. Table 7, summarizes the objective quality test of the new encoder compared to the standard EVRC coder under channel error conditions. We conducted a test for 5048-frame segment of speech (101 sec) on PII (366 MHz) computer. The EVRC algorithm lasts 107 sec to perform the encoding and decoding operations, while the new algorithm lasts only 81 sec to complete the same functions.

6. Conclusion

In this paper we proposed a new speech encoder. This encoder is a variable rate coder operates with different modes from 7.6 down to 0.8 kbits/s. The proposed encoder is a high quality with a reduced complexity. The search procedure is faster than that of standard EVRC, and the processing time is reduced by more than 24% with a better quality than the standard method. This encoder is tested with three different languages; English, Chinese, and Arabic under error free conditions and channel error conditions. The performance results show that it is robust with channel errors.

References

- [1] P. Kroon and M. Recchione, "A low-Complexity Toll-Quality Variable Bit Rate Coder for CDMA Cellular Systems", in *Proc. IEEE Int. Conf. Acoust, Speech, and Signal Processing*, 1995, pp. 5-8.
- [2] K. Mano, T. Moriya, et al., "Design of a Pitch Synchronous Innovation CELP Coder for Mobile Communications", in *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 1, January 1995, pp. 31-41.
- [3] M.R. Schroeder, and B. S. Atal, "Code-Excited Linear Prediction (CELP): high- quality speech at very low bit rates", *Proc. IEEE Int. Conf. ASSP*, 1985, pp. 937-940.
- [4] P. Kroon and W. B. Kleijn, "Linear-prediction based analysis-by-synthesis coding", in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), Amsterdam: Elsevier Science Publishers, 1995.
- [5] W. B. Kleijn, P. Kroon, L. Cellario, and D. Sereno, "A 5.85 kb/s CELP Algorithm for Cellular applications", in *Proc. Int. Conf. Acoust, Speech, Signal Processing*, 1993, pp. 596-599.
- [6] W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP Speech-Coding Algorithm", *European Transactions on Telecommunications*, Vol. 5, No. 5, Sept/Oct. 1994, pp. 573-582.
- [7] D. Nahumi, and W. B. Kleijn, "An Improved 8 kb/s RCELP Coder", *IEEE Workshop on Speech Coding*, 1995, pp. 39-40.
- [8] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders", *IEEE Trans. Speech and Audio Process*, vol. 2, no. 1, 1994, pp. 42-54.
- [9] Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems, IS-127, July 19, 1996.
- [10] R. Salami, C. Laflamme, J-P. Adoul, et al., "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech coder", *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 2, March 1998, pp. 116-130.

Table 1. New AMR Encoder Bit Allocations

Parameter	Rate 1	Rate 2	Rate 3	Rate 4
LSP	28	28	22	8
Pitch Delay	7	7	7	-
Delta Delay	5	-	-	-
ACB Gain	4×3	4×3	4×3	-
FCB Shape	4×20	4×14	4×10	-
FCB Gain	4×5	4×5	4×4	-
Frame Energy	-	-	-	8
Total bits	152	123	97	16

Table 3. Rate 1 and 2 Algebraic Codebook Structure

Track	Pulse Positions
T0	0, 5, 10, 15, 20, 25, 30, 35
T1	1, 6, 11, 16, 21, 26, 31, 36
T2	2, 7, 12, 17, 22, 27, 32, 37
T3	3, 8, 13, 18, 23, 28, 33, 38
T4	4, 9, 14, 19, 24, 29, 34, 39

Table 5. Bits Classifications for Rate 1.

Class 1	
LSP Index	28
Pitch Delay	7
Delta Delay	5
Total	40
Class 2	
ACB Gain	12
FCB Gain	20
Total	32
Class 3	
FCB Index	60
FCB Sign	20
Total	80

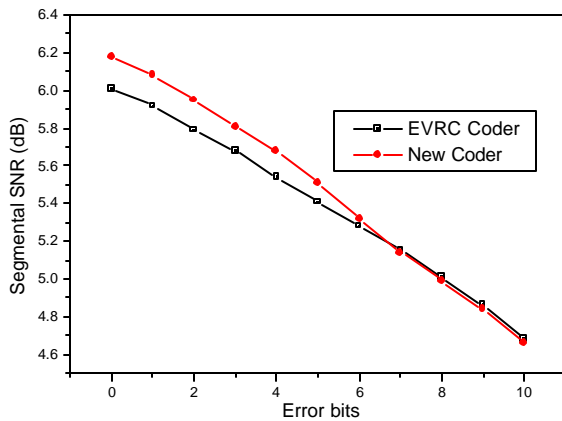


Fig. 1: Segmental SNR as a function of Error bits.

Table 2. LSP Interpolation Constants

k	segment start sample	Segment end sample	μ_k
0	0	79	0.0
1	80	119	1/6
2	120	159	1/2
3	160	199	1/2
4	200	239	5/6
5	240	319	1.0

Table 4. Rate 3 Algebraic Codebook Structure

Track	Pulse Positions
T0	0,5,10,15,20,25,30,35
T1	2,7,12,17,22,27,32,37
T2	4,9,14,19,24,29,34,39

Table 6. Objective Test under error free conditions

Data-Base	Language	English	Chinese	Arabic
	Speaker	Male	Male	Male
	Length (sec)	81	111	101
SEG SNR (dB)	New Coder	6.24	6.57	6.84
	EVRC	5.43	6.01	5.96

Table 7. Objective Test under channel error conditions

BER	EVRC	New Coder
	SEGSNR: of AWGN Channel	
2%	5.62	6.64
0.7%	5.86	6.83
SEGSNR: of Rayleigh Channel		
2%	5.51	6.54
0.7%	5.82	6.78

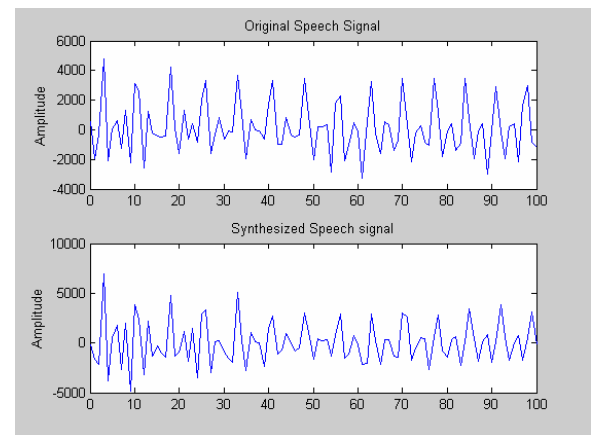


Fig.2: Original and Synth. Speech for the New System.

Biographies

Ahmed J. Jameel was born in Baghdad, Iraq on July 19, 1963. He received the B.Sc. and M.Sc. degrees from Baghdad University, Department of Electrical Engineering in 1985 and 1994, respectively.

Currently, he is a Ph.D. candidate at the Department of Radio Engineering, Southeast University. His research interests include digital communications, error-correcting codes, and speech coding for mobile communications.

You Xiaohu (S'82–M'87–SM'98) received the B.S. and M.S. degrees in electrical engineering from Nanjing Institute of Technology, Nanjing, China, in 1982 and 1985, respectively, and the Ph.D. degree (with honors), also in electrical engineering from Southeast University, Nanjing, in 1988. From 1987 to 1989, he was with Nanjing Institute of Technology as a Lecturer. From 1990 to the present time, he has been with Southeast University, first as an Associate Professor and later as a Professor. His research interests include mobile communications, adaptive signal processing, and artificial neural networks with applications to communications and biomedical engineering. He is the Chief of the Technical Group of China 3G Mobile Communication R & D Project.

Dr. You received the excellent paper prize from the China Institute of Communications in 1987 and the Elite Outstanding Young Teacher Awards from Southeast University in 1990, 1991, and 1993. He was also a recipient of the 1989 Young Teacher Award of Fok Ying Tung Education Foundation, State Education Commission of China.

Wang Ling received the B.S. in 1985 and obtained M.S degree in 1988 in Radio Engineering from the Southeast University, Nanjing, P.R.CHINA.

From 1988 to 1997, she was a teacher in Electronic Engineering Department in Southeast university and performed research and development in electronic equipments, since 1997, she has been working on digital cellular communication system with emphasis on speech coding and decoding.

Gao Xiqi received his Ph.D. degree from Southeast University in 1997. Now, he is a professor at the Department of Radio Engineering, Southeast University. His main interests include mobile access techniques, space-time signal processing for mobile communications, wavelets and filter banks.