DPCM with Forward Gain-Adaptive Quantizer and Simple Switched Predictor for High Quality Speech Signals

Vladimir M. DESPOTOVIC[1], Zoran H. PERIC[2], Lazar VELIMIROVIC[2] and Vlado D. DELIC[3]
[1]*University of Belgrade, Technical Faculty of Bor, Bor, 19210, Serbia*
[2]*University of Nis, Faculty of Electronic Engineering, Nis, 18000,Serbia*
[3]*University of Novi Sad, Faculty of Technical Sciences, Novi Sad, 21000, Serbia*
vdespotovic@tf.bor.ac.rs

***Abstract*—In this article DPCM (Differential Pulse Code Modulation) speech coding scheme with a simple switched first order predictor is presented. Adaptation of the quantizer to the signal variance is performed for each particular frame. Each frame is classified as high or low correlated, based on the value of the correlation coefficient, then the selection of the appropriate predictor coefficient and bitrate is performed. Low correlated frames are encoded with a higher bitrate, while high correlated frames are encoded with a lower bitrate without the objectionable loss in quality. Theoretical model and experimental results are provided for the proposed algorithm.**

***Index Terms*—adaptive quantization, companding, correlation, speech processing, prediction**

I. INTRODUCTION

This paper presents a simple, yet very effective middle bitrate, high quality speech coding scheme based on Differential Pulse Code Modulation (DPCM) and correlation. DPCM encodes the difference between sample points to compress the digital data. Since audio waves propagate in predictable patterns, DPCM predicts the next sample and encodes the difference between the predicted and the actual point. The differences are smaller numbers than the numerical value of each sample and thereby reduce the resulting bitstream.

Since speech is a non-stationary process and tends to change in time [1-3], in order to process speech effectively it is necessary to segment speech waveform into frames. Typically, frame is selected between 10 and 30 ms [4-6]. Correlation coefficient is determined for each frame and frame is classified as high or low correlated. Low correlated frames are encoded with a higher bitrate (i.e. 7 bits/sample), while high correlated frames are encoded with a lower bitrate (i.e. 6 bits/sample) without loss in quality of the reconstructed speech signal. Switched first order predictor is used in DPCM scheme, with two possible fixed values of predictor coefficients: for low (close to zero), and for high correlated samples (close to one).

Forward gain-adaptive quantizer based on optimal companding model presented in [7] is used for quantization of difference between samples in DPCM. Companding algorithms reduce the dynamic range of an audio signal, thus reducing the quantization error. This can be traded for the reduced bit rate for equivalent quality of speech. For instance, widely used ITU-T G.712 standard for audio companding preserves high quality of speech, while converting digital, linear 12 bit signal into 8-bit code [8-9].

It will be shown that the presented speech coding scheme effectively exploits correlation between samples, giving a substantially lower bitrate even for very low prediction orders, for the same quality of speech.

The remainder of the article is organized as follows. Section II describes the theoretical basics of the optimal quantizer design. Section III presents concrete realization of DPCM with gain-adaptive quantizer and the first order predictor. Section IV discusses the theoretical model and experimental results. Finally, section V gives a conclusion.

II. THEORETICAL BACKGROUND AND QUANTIZER DESIGN

Consider an N-point scalar quantizer Q_N designed optimally in the minimum mean square sense for the probability density function (pdf) $q(x)$ that is applied to a source with pdf $p(x)$ where:

$$p(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}}{\sigma} |x| \right) \tag{1}$$

$$q(x) = \frac{1}{\sqrt{2}\hat{\sigma}} \exp\left(-\frac{\sqrt{2}}{\hat{\sigma}} |x| \right) \tag{2}$$

Laplacian distribution with zero mean is assumed, having in mind the fact that the speech signal distributions are not Gaussian, but rather Laplacian or Gamma based [10].

Q_N is characterized by a set of N representation levels $y_1 < y_2 < ... < y_N$ and decision thresholds $t_0 < t_1 < ... < t_N$ with $t_0 = -\infty$ and $t_N = \infty$ [11]. If we assume that a total number of reconstruction points N is large enough, distortion can be well approximated by Bennett's integral [12]:

$$D \approx \frac{1}{12N^2} \int_{-\infty}^{\infty} \frac{p(x)}{\lambda^2(x)} dx \tag{3}$$

where $\lambda(x)$ denotes asymptotically optimal point density of Q_N [11]:

$$\lambda(x) = \frac{q^{\frac{1}{3}}(x)}{\int_{-\infty}^{\infty} q^{\frac{1}{3}}(x) dx} = \frac{1}{3\sqrt{2}\hat{\sigma}} \exp\left(-\frac{\sqrt{2}}{3\hat{\sigma}} |x| \right) \tag{4}$$

Digital Object Identifier 10.4316/AECE.2010.04015

Substituting $p(x)$ and $\lambda(x)$ in (3), an expression for distortion is obtained:

$$D = \frac{9\hat{\sigma}^2}{2N^2C} \tag{5}$$

where $C = 3 - \frac{2\sigma}{\hat{\sigma}}$. In the case of matched quantization, i.e. $\sigma = \hat{\sigma}$, the above reduces to Panter-Dite formula $D = \frac{9\hat{\sigma}^2}{2N^2}$ [13]. Signal to quantization noise ratio can be determined as:

$$SQNR = 10\log_{10}\left(\frac{\sigma^2}{D}\right) = 6.02R + 10\log_{10}\frac{C(3-C)^2}{18} \tag{6}$$

where $N = 2^R$ and R is a bit rate.

The voiced and unvoiced speech signals can be classified by a correlation coefficient [14]. Due to the concentration of low frequency energy of voiced sounds, the adjacent samples of voiced speech are highly correlated, with correlation coefficient close to one [15]. On the other hand, the correlation is close to zero for unvoiced speech. Taking into account the correlation between speech samples, the variance of the quantized signal becomes $\hat{\sigma}^2(1-\rho^2)$, so distortion can be determined as:

$$D_{corr} = \frac{9\hat{\sigma}^2}{2N^2C}(1-\rho^2) \tag{7}$$

where ρ is the correlation coefficient. Notice that the correlation coefficient is a number between -1 and 1, which indicates the degree of the linear dependence between the speech samples. The correlation coefficient is +1 in the case of a perfect positive (increasing) linear relationship, −1 in the case of a perfect decreasing (negative) linear relationship [16], and some value between −1 and 1 in all other cases. As it approaches zero, there is less of a correlation. Substituting equation (7) in the expression for SQNR we obtain:

$$SQNR_{corr} = 6.02R + 10\log_{10}\frac{C(3-C)^2}{18(1-\rho^2)} = SQNR + G_p \tag{8}$$

where $G_p = 10\log_{10}\frac{1}{1-\rho^2}$ is correlation (or prediction) gain, which is dependent only on the correlation coefficient.

Knowing that the percentage of silence in speech is normally around 25% [17], speech waveform should be first divided into frames before further processing, then each frame should be classified as high correlated (voiced speech) or low correlated (unvoiced speech or silence), and finally weighting of SQNR should be performed according to this classification:

$$SQNR_{total} = w \cdot SQNR_{corr}^{(1)} + (1-w) \cdot SQNR_{corr}^{(2)} \tag{9}$$

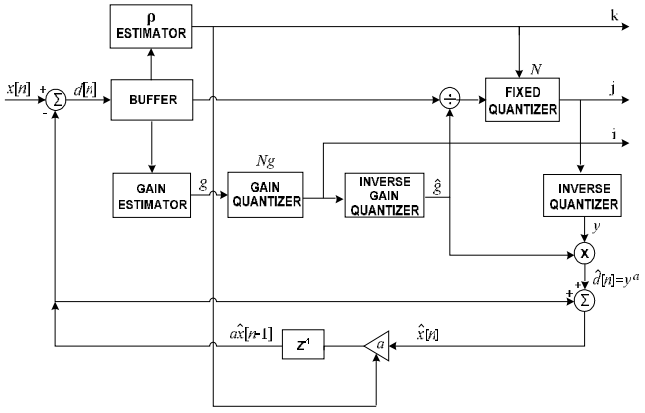
where $w = 0.75$ and

$$SQNR_{corr}^{(i)} = 6.02R_i + 10\log_{10}\frac{C(3-C)^2}{18(1-\rho_i^2)}, \quad i = 1,2 \tag{10}$$

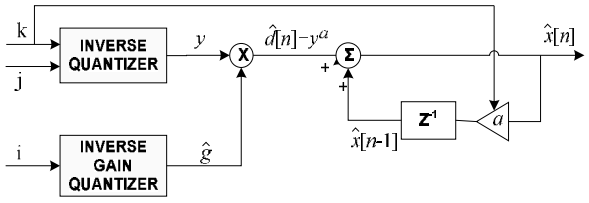
The lower bit rate can be used for quantization of voiced speech segments (i.e. R_1 = 6 bits/sample) with the higher correlation coefficient (e.g. $\rho_1 = 0.8$), while segments of silence (or unvoiced speech) can be encoded with the higher bit rate (i.e. R_2 = 7 bits/sample) and with the lower correlation coefficient (e.g. $\rho_1 = 0.3$).

III. DPCM WITH GAIN-ADAPTIVE QUANTIZER AND SIMPLE PREDICTOR

A simple but useful DPCM scheme with a first order predictor will be used. This coder has a predictor adaptation strategy that depends only on the previous codeword. Adaptation of the quantizer will be performed using the forward gain-adaptive quantizer based on the optimal companding model presented in [7], as shown in Fig. 1. The quantizer is adapted in response to a short-term estimate of the input signal standard deviation $\hat{\sigma}_n$. This may be achieved by scaling all the samples by a gain factor $\hat{g} = \hat{\sigma}_n$, however it is preferably from a complexity standpoint to divide the input to the quantizer by the estimated gain.



(a)



(b)

Fig. 1. (a) Encoder; (b) Decoder

Consider a speech coding scheme consisted of a buffer, correlation estimator, non-adaptive N-level scalar quantizer, gain estimator and N_g-level scalar quantizer (for gain quantization), as shown in Fig. 1(a). Buffer is used for division of input speech signal into frames. The average correlation coefficient ρ is then calculated for each frame:

$$\rho = \frac{\sum_{i=1}^{M-1} x_i \cdot x_{i+1}}{\sum_{i=1}^{M-1} x_i^2} \tag{11}$$

where M is the length of the frame. If $\rho < 0.6$ input signal to the quantizer is considered low correlated and will be

encoded with 7 bits/sample. Vice-versa, if $\rho \geq 0.6$ the lower bit rate 6 bits/sample will be used for a high correlated signal. At the same time, the switched first order predictor will be used, where the predictor coefficients will be chosen according to the value of the estimated correlation coefficient (close to zero for the low correlated, and close to one for the high correlated frames). The main idea of the proposed coding scheme is that for the high enough correlation, coding with less bits/sample is possible without the objectionable loss in a quality of the reconstructed speech. If the number of frames that satisfy this condition is significant, the saving in bit rate can be substantial.

Gain-adaptive quantizer based on the companding model presented in [7] is realized by scaling the codebook of the following fixed quantizer:

$$t_j = \frac{3}{\sqrt{2}} \ln\left(\frac{N}{2N - 2j + (2j - N)\exp\left(-\frac{\sqrt{2}}{3} A_{\max} \right)} \right) \quad (12)$$

$$y_j = \frac{3}{\sqrt{2}} \ln\left(\frac{N}{2N - 2j + 1 + (2j - 1 - N)\exp\left(-\frac{\sqrt{2}}{3} A_{\max} \right)} \right) \quad (13)$$

with the value of the estimated standard deviation (gain). Notice that the symmetry of the decision thresholds (t_j) and the representation levels (y_j) is assumed, i.e. $N/2 \leq j < N$ in equation (13). $A_{\max}$ is the maximum amplitude of the input signal to the quantizer, and can be determined as [7]:

$$A_{\max} = \frac{3}{\sqrt{2}} \ln\left(\frac{N}{2} \right) + \sqrt{2} \quad (14)$$

N is the number of the representation levels and it is equal to 64 (6 bits/sample) for the high correlated frames, or 128 (7 bits/sample) for the low correlated frames.

It is necessary to modify the codebook to account for the gain normalization. After scaling, the decision thresholds and the representation levels of the gain-adaptive quantizer, denoted with t_j^a and y_j^a respectively, are determined as:

$$t_j^a = \hat{g} \cdot t_j \quad \text{and} \quad y_j^a = \hat{g} \cdot y_j, \quad j = 1,2,...,N \quad (15)$$

where $\hat{g}$ is estimated gain (standard deviation).

$$\hat{g} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} d_i^{\,2}} \quad (16)$$

Finally, the information about the gain needs to be transferred to the decoder as a side information, so it needs to be quantized as well using N_g quantization levels. The representation levels for the estimated gain are determined using the logarithmic quantizer from:

$$20\log_{10}(\hat{g}_i) = 20\log_{10}(\sigma_{\min}) + (2i-1)\frac{\Delta}{2}, \quad i = 1,2,...,N_g \quad (17)$$

where $\Delta = \dfrac{20\log_{10}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}{N_g}$ and dynamic range of the input signal power is defined by $\left[20\log_{10}(\sigma_{\min}), 20\log_{10}(\sigma_{\max})\right]$. The reconstructed speech signal $\hat{x}$ will be determined as:

$$\hat{x}[n] = a \cdot \hat{x}[n-1] + y^a[n] \quad (18)$$

where n denotes the n-th sample of the input speech signal and y^a the output of the adaptive quantizer. Notice that switched first order predictor is used, with predictor coefficient a close to zero for low (e.g. $a = 0.3$) and close to one for high correlated frames (e.g. $a = 0.8$)

IV. RESULTS AND DISCUSSION

Theoretical results are given for a model presented in Section II (see equations (9) and (10)), and compared with cases when speech is encoded with constant bitrates, 6 or 7 bits/sample (equation (6) with R=6 or R=7), as shown in Fig. 2. Cases when 75% (dashed line) and 60% (solid line) of speech is high correlated (voiced speech) are analyzed. High correlated signal is encoded with bitrate 6 bits/sample, while low correlated signal is encoded with 7 bits/sample. *SQNR* dependence on the signal variance (-20dB÷20dB dynamic range is assumed) for the 16-levels gain quantization is given in Fig. 2, i.e. we have a case of the switched quantizer designed optimally for 16 different values of the variance (gain) in order to cover the whole dynamic range of the input speech signal. Notice that for the optimal (matched) quantizer ($\sigma = \hat{\sigma}$) *SQNR* has a maximum value. It is obvious that the proposed model satisfies ITU-T G.712 standard at any point.

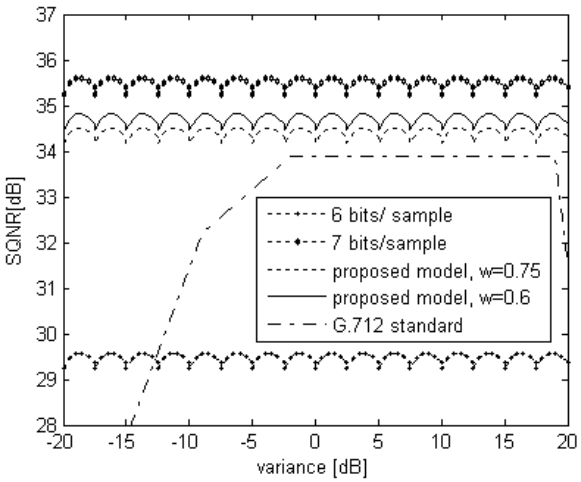


Fig. 2. Theoretical results: SQNR vs. signal variance

The properties of the proposed model are given in Table I, where ρ_1 denotes the correlation coefficient of high, and ρ_2 of low correlated speech. Results in Fig. 2 and Table I show that although bitrate for the proposed model is significantly lower than 7 bits/sample, loss in SQNR in both cases is less than 1 dB.

TABLE I. PROPERTIES OF THE PROPOSED THEORETICAL MODEL

w	ρ_1 (high correlated)	ρ_2 (low correlated)	Bitrate [bits/sample]
0.75	0.8	0.3	6.25
0.60	0.8	0.3	6.40

The experimental results are provided using the speech coding scheme given in Section III. All the experiments are performed on 3 minutes of speech extracted from the TIMIT database (56 separate sentences, 31 male and 25 female American English speakers). Speech is divided into frames with the frame length M=200.

TABLE II. EXPERIMENTAL RESULTS
AVERAGE $SQNR_{SEG}$ AND BITRATE

***Ng*-level quantizer**	**$SQNR_{seg}$ [dB]**	**Bitrate [bits/sample]**
2 (1bit/frame)	19.7082	6.340
4 (2bits/frame)	30.0016	6.345
8 (3 bits/frame)	34.0595	6.350
16 (4 bits/frame)	34.5972	6.355
32 (5 bits/frame)	34.6945	6.360

Results of the segmental signal to noise ratio ($SQNR_{seg}$) and bitrate are given in Table II for different number of gain quantization levels: $N_g = 2$, $N_g = 4$, $N_g = 8$, $N_g = 16$ and $N_g = 32$. $SQNR_{seg}$ was used as a measure of quality of speech. It is considered a better perceptual model compared to the traditional signal to quantization noise ratio since it evaluates the quantization noise with the respect to the energy in each underlying speech frame. Information about the gain and the estimated correlation coefficient are transferred as side information to decoder. This slightly increases the necessary bitrate (1 bit/frame for the correlation coefficient and 1-5 bits/frame for the gain). The experimental results for the number of gain quantization levels 8 to 32 show very good match with the theoretical model. Value of the average correlation coefficient $\rho = 0.6$ was used as a threshold for classification of speech as low or high correlated. The experimental results have shown that 67% of samples were high correlated, while 33% were classified as low correlated using this criterion. This is also in accordance with the other results of speech/silence classification [17].

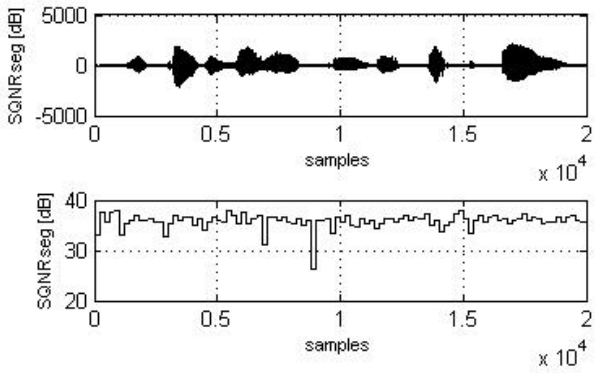


Fig. 3. Experimental results: $SQNR_{seg}$ over different frames of speech

Fig. 3. shows signal to noise ratio over the different segments of input speech signal for one selected sentence (female speaker, 2.5 seconds duration of speech).

V. CONCLUSION

In this paper a simple DPCM speech coding scheme with the first order switched predictor and gain-adaptive quantizer is presented. Quantizer is realized using the companding model and it is adapted to a short-term estimate of the input signal standard deviation on each frame. Each frame is first classified as low or high correlated based on the value of the correlation coefficient, and than encoded with the corresponding bitrate (lower for high correlated, and higher for low correlated frames). The main idea is that it is possible to encode speech signal with less bits/sample without the objectionable loss in performance, if a high enough correlation exists in a frame. Savings of up to 0.75 bits/sample are obtained with loss in SQNR of only 1 dB. The ITU-T G.712 standard is also satisfied in a whole dynamic range of the input signal powers. We provide theoretical model and experimental results that are very well matched.

REFERENCES

[1] A. Gersho, R. M. Gray, Vector Quantization and Signal Compression, Kluwer, Academ. Pub., Chapters 5–6, pp. 133–202, 1992.
[2] N. S. Jayant, P. Noll, Digital Coding of Waveforms, Prentice-Hall, New Jersey, Chapter 4, pp. 129–139, 1984.
[3] A. Kondoz, Digital Speech, Coding for Low Bit Rate Communication Systems, JohnWiley & Sons, New Jersey, 2004.
[4] W.C. Chu, Speech Coding Algorithms, John Wiley & Sons, New Jersey, Chapters 5–6, pp. 143–183, 2006.
[5] D. Minoli, Voice over MPLS – Planning and Designing Networks, McGraw-Hill, Chapters 1–2, pp. 1–134, 2002.
[6] O. Hersent, J. Petit, D. Gurle, Beyond VoIP Protocols – Understanding Voice Technology and Networking Techniques for IP Telephony, John Wiley & Sons, New Jersey, Chapters 1–2, pp. 1–88, 2005.
[7] J. Nikolic, Z. Peric, "Lloyd–Max's Algorithm Implementation in Speech Coding Algorithm Based on Forward Adaptive Technique", *Informatica*, vol. 19, (2), pp. 255–270, 2008.
[8] B. Khasnabish, Implementing Voice over IP, John Wiley & Sons, New Jersey, 2003.
[9] ITU-T Recommendation G.712: Transmission Performance Characteristics of Pulse Code Modulation (PCM), 1992.
[10] S. Gazor, W. Zhang, "Speech probability distribution", *IEEE Signal Processing Letters*, vol. 10 (7), pp. 204-207, 2003.
[11] S. Na, "Asymptotic Formulas for Mismatched Fixed-Rate Minimum MSE Laplacian Quantizers", *IEEE Signal Processing Letters*, vol. 15, pp. 13–16, 2008.
[12] W. R. Bennett, "Spectra of quantized signals", *Bell Syst Tech J*, vol. 27, pp. 446–472, 1948.
[13] P. F. Panter, W. Dite, "Quantization distortion in pulse count modulation with nonuniform spacing of levels", in *Proc. IRE*, vol. 39 (1), pp. 44-48, 1951.
[14] J. R. Deller, J. H. L. Hansen, J. G. Proakis, Discrete-Time Processing of Speech Signals, IEEE Press, 2000.
[15] C. C. Cho, N. I. Park, H. K. Kim, "A Packet Loss Concealment Algorithm Robust to Burst Packet Loss for CELP-type Speech Coders", in *Proc. of ITC-CSCC 2008*, pp. 941-944, 2008.
[16] S. Dowdy, S. Wearden, Statistics for Research, John Wiley and Sons, New York, 1983.
[17] R. Goldberg, L. Riek, A Practical Handbook of Speech Coders, CRC Press, 1. edition, 2000.