Multilevel Delta Modulation with Switched First-Order Prediction for Wideband Speech Coding

Zoran Peric[1], Bojan Denic[1], Vladimir Despotovic[2]
[1]*Faculty of Electronic Engineering, University of Nis,*
Aleksandra Medvedeva 14, 18000 Nis, Serbia
[2]*Technical Faculty in Bor, University of Belgrade,*
Vojske Jugoslavije 12, 19210 Bor, Serbia
bojan.denic@elfak.rs

***Abstract*—In this paper a delta modulation speech coding scheme based on the ITU-T G.711 standard and the switched first-order predictor is presented. The forward adaptive scheme is used, where the adaptation to the signal variance is performed on frame-by-frame basis. The classification of the frames into weakly and highly correlated was done based on the correlation coefficient calculated for each frame, providing a basis for choosing the appropriate predictor coefficient. The obtained results indicate that the proposed model significantly outperforms the scalar companding system based on the G.711 standard. The obtained experimental results were verified using the theoretical model in the wide dynamic range of the input variance.**

***Index Terms*—Quantization; Delta modulation; Correlation coefficient; Speech coding; Signal to noise ratio.**

I. INTRODUCTION

Speech coding is the process of obtaining a compact representation of speech signals for efficient transmission and/or storing in digital media [1]–[3]. Speech coders are essential parts of Public Switched Telephone Networks (PSTN), Voice over Internet Protocol (VoIP), mobile communications, videoconferencing etc. To this end it is fundamental to discover the speech coding algorithms that provide high intelligibility and quality of speech at the consumer side.

The high bit rate ITU-T G.711 codec [4] and its extensions G.711.0 and G.711.1 have been accepted as a standard in many modern speech coding applications. G.711 is the companding system employing the piecewise linear approximation to the μ-law or the A-law logarithmic compression function. Its main qualities are low complex encoding algorithm and small delay for the high-quality reproduced speech [4]. The wideband extension of G.711 known as G.711.1 is proposed in [5], [6] and has been standardized for wideband audio and speech signal processing. The authors in [7] proposed the two-stage quantization with embedded G.711 coder in the first processing stage, followed by the segmental uniform quantization that performs the reduction of the quantization error introduced in the first stage. In this way higher signal quality is achieved in comparison with the G.711 quantizer.

As a nonstationary process speech is usually transmitted in frames (certain number of samples), since the speech properties remain mostly unchanged within one frame. To obtain the desired performance, the quantizer requires some kind of adaptation at the frame level, e.g. using the variance or the probability density function (pdf). Different types of adaptation have been used in speech coding algorithms based on Pulse Code Modulation [8], [9], Differential Pulse Code Modulation (DPCM) [10] or Delta Modulation (ΔM) [11].

The later approaches, DPCM and ΔM as its special case, belong to the class of predictive coding algorithms [1]–[3], [12], [13]. Delta modulation has become an attractive method for signal processing, due to its simple architecture [14]. In particular, it includes one-bit quantizer along with the first-order predictor. Various modifications of ΔM have been proposed over the years to improve the performance of the basic structure, including the adaptive delta modulation [11], and the sigma-delta modulation.

High-quality DPCM speech coding scheme employing the scalar companding quantizer and the switched first-order predictor is proposed in [10]. In this paper we keep the switched first-order predictor, but propose the modified ΔM configuration, with the embedded high-rate G.711 quantizer, and denote it as the multilevel delta modulation system. The adaptation to the signal statistics is performed frame-wise using the short-term estimate of the variance. The forward adaptive scheme was used, as it offers better performance with respect to the backward adaptation [15], and it is less sensitive to the transmission error [1]; however it requires sending the side information to the decoder side. The switched predictor chooses between two coefficients, one for weakly and one for highly correlated frames, based on the correlation coefficient calculated for the particular frame.

We test the performance of the proposed algorithm in the real environment using the speech signal, and we use Signal

Manuscript received 12 May, 2017; accepted 28 September, 2017.
This work was supported in part by the by the Ministry of Education and Science of the Republic of Serbia, grant no. TR 32051 and TR 32035, within the Technological Development Program.

to Noise Ratio (*SNR*) as a measure of performance. The efficiency of the proposed algorithm is compared to the scalar companding system based on the G.711 standard [4].

The rest of the paper is organized as follows: Section II describes the companding system with embedded G.711 quantizer. Section III gives the overview of the proposed multilevel delta modulation speech coding scheme. Section IV summarizes and discusses the experimental results, and finally Section V concludes the paper.

II. Scalar Companding System Based on the G.711 Recommendation

A. Non-adaptive Scalar Compandor

In the companding quantization, the input is first transformed using a nonlinear compressor function, then further quantized using the uniform quantizer, and finally restored in expander using the inverse nonlinear function. The compressor and expander form a compander. In the G.711 standard [4], the piecewise linear approximation to the logarithmic μ-law compression function is performed, which is given by

$$c(x) = \frac{x_{\max}}{\ln(1+\mu)} \ln\left(1 + \frac{\mu|x|}{x_{\max}}\right) \operatorname{sgn}(x), \tag{1}$$

where $|x| \le x_{\max}$, μ is the compression factor and $x_{\max}$ is the upper support region threshold.

The support region $[-x_{\max}, x_{\max}]$ of the quantizer is divided into $2L$ segments (L positive and L negative), where each segment is composed of m uniform cells. Each consecutive segment in the positive part of the quantizer characteristic is twice as large as the previous. As the quantizer characteristic is symmetric, the same holds true for the negative segments. The segments width denoted by Δ_i are determined as

$$\Delta_i = 2^i \frac{x_{\max}}{255m}, \tag{2}$$

where $i = 0,1,...,L-1$, and $x_{\max}$ is given as in [16]

$$x_{\max} = \frac{1}{\sqrt{2}} \log\left(\frac{3\mu N^2}{\log(\mu+1)}\right), \tag{3}$$

where N is the number of quantization levels.

The borders between the segments x_i are given as

$$x_i = \frac{\left(2^i - 1\right)x_{\max}}{255}, \tag{4}$$

where $i = 0,1,...,L$ while the cells borders x_{ij} and the representative levels y_{ij} in i-th (i = 0, 1,..., $L-1$) segment are given by

$$x_{ij} = x_i + j\Delta_i, \tag{5}$$

where $j = 0,1,...,m$,

$$y_{ij} = x_i + \frac{(2j-1)}{2}\Delta_i, \tag{6}$$

where $j = 1,...,m$. If we assume that information source is memoryless Laplacian with zero mean and unit variance having PDF [1]

$$p(x,\sigma) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}|x|}{\sigma}\right), \tag{7}$$

where σ is the standard deviation, then the mean-squared distortion, which is a measure of irreversible error incurred during the quantization, consists of granular D_g and overload D_0 distortion [1]:

$$D_g = \sum_{i=0}^{L-1} \frac{\Delta_i^2}{12}\left(\exp\left(-\frac{\sqrt{2}x_i}{\sigma}\right) - \exp\left(-\frac{\sqrt{2}x_{i+1}}{\sigma}\right)\right), \tag{8}$$

$$D_o = \exp\left(-\frac{\sqrt{2}x_{\max}}{\sigma}\right) \times \\ \times \left(\left(x_{\max} - y_{L-1,m} + \frac{\sigma}{\sqrt{2}}\right)^2 + \left(\frac{\sigma}{\sqrt{2}}\right)^2\right), \tag{9}$$

where $y_{L-1,m}$ is the representative level of the last cell in the last segment, that can be determined from (6).

Along with distortion, we use Signal to Quantization Noise Ratio for the performance estimation [1]–[3]

$$SQNR = 10\log_{10}\left(\frac{\sigma^2}{D_g + D_o}\right). \tag{10}$$

B. Adaptive Scalar Compandor

Forward adaptive coding scheme operating on frame-by-frame basis is illustrated in Fig. 1, where the adaptation to the short-term estimate of the variance is done for each frame of input signal. The building blocks of such scheme are a buffer, a variance estimator, a log-uniform quantizer Q_{LU} with L levels for the quantization of frame variance and an adaptive scalar compandor (Q), which codebook is updated frame-wise.

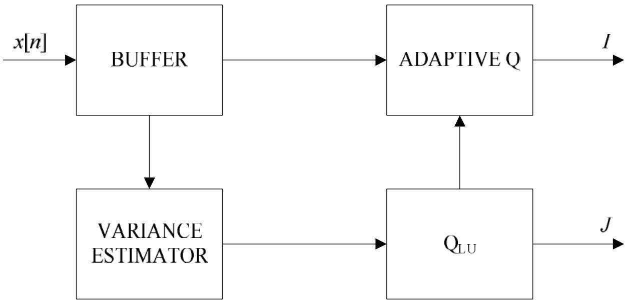


Fig. 1. Forward adaptive coding scheme.

It works in the following way. The buffer stores one frame or M samples of the input signal and the variance of the input speech is determined in the variance estimator

$$\sigma_x^2 = \frac{1}{M}\sum_{i=1}^{M} x_i^2. \tag{11}$$

For the variance quantization we use log-uniform quantizer having L-levels which outputs are

$$20\log_{10}(g_i) = 20\log_{10}(\sigma_{\min}) + \frac{(2i-1)}{2}\Delta, \tag{12}$$

where $i = 1,...,L$, $\Delta = 20\log_{10}(\sigma_{\max}/\sigma_{\min})/L$ and dynamic range of the input signal is defined as [20 $\log_{10}(\sigma_{\min})$, 20 $\log_{10}(\sigma_{\max})$].

The quantizer codebook is updated according to the quantized variance, hence, for adaptive threshold and levels we get: $t_a = \sqrt{g_i} \times t_f$ and $y_a = \sqrt{g_i} \times y_f$ where t_f and y_f are threshold and levels of nonadaptive quantizer, respectively.

Observe in Fig. 1 two digital signals I and J, where signal I carries $\log_2 N$ bit code-words that represent signal within the frame and signal J carries the information for quantized variance used for adaptation consisted of $\log_2 L$ bits per frame (additional or side information).

The bit rate of the forward adaptive quantizer is given by

$$R_{\text{PCM}} = \log_2 N + \frac{R_{\text{LU}}}{M}, \tag{13}$$

where $R_{\text{LU}} = \log_2 L$ bits is side information.

In order to provide the appropriate theoretical analysis in a wide dynamic range of the input variances, we define the distortion and *SQNR* for the particular variance:

$$D(\sigma_i) = \sum_{i=1}^{N}\int_{t_{ai-1}}^{t_{ai}} (x - y_{ai})^2 p(x,\sigma_i)\,dx, \tag{14}$$

$$SQNR(\sigma_i) = 10\log_{10}\left(\frac{\sigma_i^2}{D(\sigma_i)}\right). \tag{15}$$

III. DELTA MODULATION WITH ADAPTIVE G.711 QUANTIZER AND SWITCHED FIRST-ORDER PREDICTOR

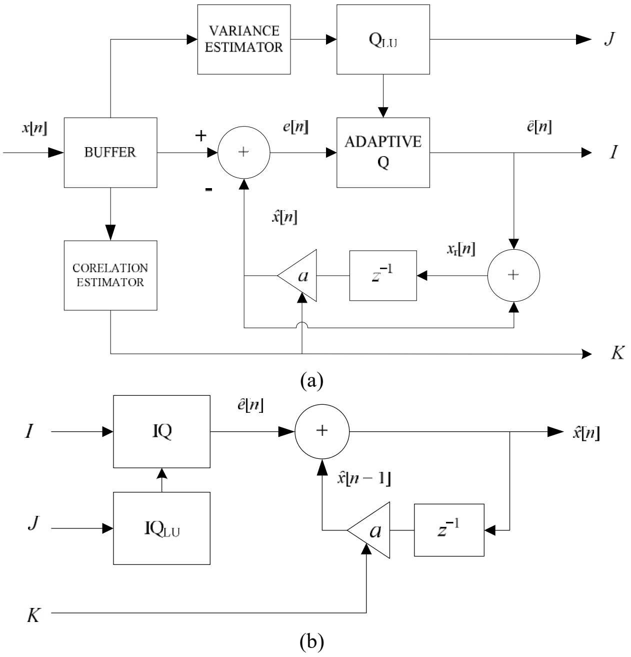
A simple delta modulation scheme with a switched first-order predictor is depicted in Fig. 2, where the adaptive scalar compandor is implemented using the forward adaptive scheme described in previous section.

The switched predictor in the feedback has at disposal two coefficients a_1 and a_2, and it chooses one of them according to correlation coefficient ρ estimated for each frame

$$\rho = \frac{\sum_{i=1}^{M-1} x_i x_{i+1}}{\sum_{i=1}^{M} x_i^2}. \tag{16}$$

Specifically, if $\rho < 0.8$ the input frame is classified as weakly correlated and the switched predictor uses the coefficient a_1, otherwise the frame is considered as highly correlated and the coefficient a_2 is employed.

The introduced coding scheme works in a similar manner as described in Section II, where the prediction error signal $e[n] = x[n] - \hat{x}[n]$ is fed to the quantizer input, where $x[n]$ is the original sample value and $\hat{x}[n]$ is the predicted sample value provided at the local decoder in the feedback of the system.



(a)

(b)

Fig. 2. Delta modulation system: (a) coder; (b) decoder.

The prediction error signal is obtained such that the first sample in each frame $x[1]$ is predicted using the last sample from the previous reconstructed frame $\hat{x}[M]$, except for the first frame where $x[1] = 0$, since there is no previous frame in that case. Hence, frames should overlap by one sample.

Note that adaptation to the variance $\sigma_e^2 = \sigma_x^2(1 - \rho^2)$ is performed for each frame, where σ_x^2 is the variance of input signal and σ_e^2 is the variance of prediction error.

Encoder (Fig. 2(a)) sends to decoder one signal more (index *K*) compared to the one in Fig. 1, since one bit information about the selected switched predictor coefficient has to be transmitted, giving the bit rate

$$R_{\Delta\text{M}} = \log_2 N + \frac{R_{\text{LU}} + 1}{M}. \tag{17}$$

Decoder (Fig. 2(b)) decodes the signal samples for each signal frame based on indices *I, J* and *K*.

In predictive coding, the overall Signal to Noise Ratio has two components, *SQNR* of the quantizer (see (15)) and the prediction gain *G* defined as [1]

$$G = 10\log_{10}\left(\frac{1}{1-\rho^2}\right). \tag{18}$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In the beginning of this section, we present the theoretical results for two variants of the scalar companding system based on G.711, non-adaptive and adaptive, described in Section II, which are used as baselines. We assume Laplacian source signal at the input in the wide dynamic range of 50 dB, and we adopt $\sigma_{\mathrm{ref}}^2 = 2 \times 10^{-3}$.

The robustness of the considered non-adaptive scalar compandor is shown in Fig. 3, where the *SQNR* is plotted as a function of input signal variance using (10).

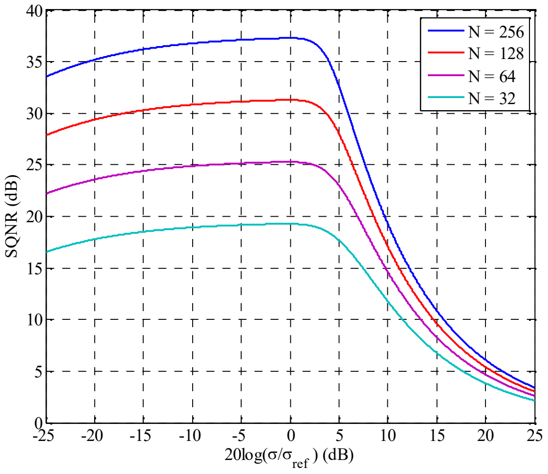


Fig. 3. Theoretical model: Non-adaptive scalar compandor (Section II-A) for various N and $\mu = 255$ in a wide dynamic range.

Figure 4 plots the theoretical *SNR* using (15) of the forward adaptive scalar compandor in the assumed variance range, for $L = 32$ levels log-uniform quantizer used for variance quantization and different number of quantization levels N for the adaptive quantizer. As it is evident, *SNR* is quite constant across the entire range.

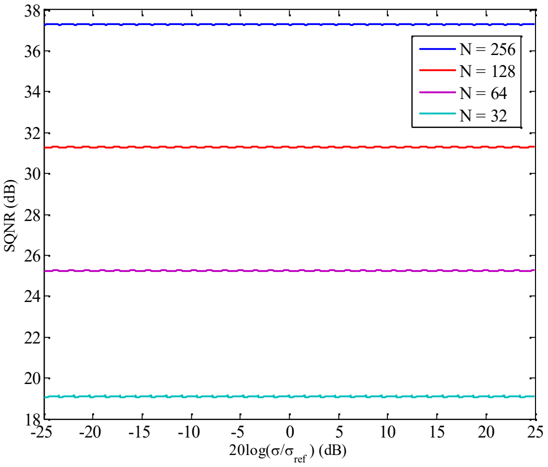


Fig. 4. Theoretical model: Adaptive scalar compandor (Section II-B) for various N, $\mu = 255$ and $L = 32$ for Q_{LU} in a wide dynamic range.

Let us further analyse the theoretical multilevel adaptive ΔM model with the switched first-order predictor, which is equivalent to the proposed coder in Section III. It is known that the percentage of silence in speech is normally around 25 % [17]; hence we adopt the weight $w = 0.25$ that defines the share of weakly correlated frames. The adjacent samples in speech signal are highly correlated with correlation coefficient close to one [1]; hence for voiced frames of speech we adopt $\rho_2 = a_2 = 0.97$. On the other hand, for weakly correlated frames we use $\rho_1 = a_1 = 0.3$. The equivalent gain of the switched predictor can be calculated as

$$G_{\mathrm{eq}} = wG_1 + (1-w)G_2, \tag{19}$$

where G_1 and G_2 refers to gain of weakly and highly correlated frames, respectively. The theoretical results (overall *SNR*) in this case are presented in Fig. 5 and show an evident improvement over the two non-predictive models of approximately 10 dB.

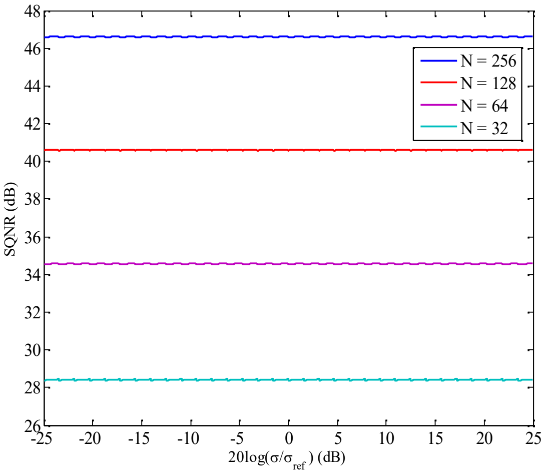


Fig. 5. Theoretical model: multilevel adaptive ΔM with the switched first order predictor for various N, $\mu = 255$ and $L = 32$ for Q_{LU} in a wide dynamic range.

Furthermore, we performed experiments on the speech signal that consists of 66 500 speech samples, sampled at 16 kHz. Speech is divided into F frames, each composed of M samples. We use $L = 32$ levels log-uniform quantizer for the frame variance quantization. As an objective measure of performance we use Signal to Noise Ratio (*SNR*).

SNR for j-th frame is determined as

$$SNR_j = 10\log_{10}\frac{\frac{1}{M}\sum_{i=1}^{M} x_{ij}^2}{\frac{1}{M}\sum_{i=1}^{M}\left(x_{ij} - \hat{x}_{ij}\right)^2}, \tag{20}$$

where x_{ij} and $\hat{x}_{ij}$ are the input and the output speech samples, respectively.

The average *SNR* is given by

$$SNR = \frac{1}{F}\sum_{j=1}^{F} SNR_j. \tag{21}$$

Let us assume that P out of F frames are classified as weakly correlated, then using (20) we have:

$$SNR_{\mathrm{wc}} = \frac{1}{P}\sum_{j=1}^{P} SNR_j, \tag{22}$$

$$SNR_{\mathrm{hc}} = \frac{1}{F-P}\sum_{j=1}^{F-P} SNR_j, \tag{23}$$

where indices “wc” and “hc” define weakly and highly correlated frames, respectively.

SNR of the whole system is given by

$$SNR = wSNR_{\mathrm{wc}} + (1-w)SNR_{\mathrm{hc}}, \tag{24}$$

where $w = P / F$ is the experimentally determined probability of occurrence of the weakly correlated frame.

In Fig. 6 we present the correlation coefficient estimated using (16) and *SNR* using (21) over all frames of size $M = 80$ for the proposed multilevel ΔM ($\mu = 255$, $N = 256$, $L = 8$, $m = 16$, $R_{\mathrm{G.711}} = \log_2 N = 8$ bit/sample) [4]. As it is obvious, in the areas of active speech the correlation coefficient ρ is close to 1, indicating the high predictability of the signal. Moreover, note higher *SNR* in the active speech area (up to 60 dB), and lower *SNR* in inactive speech frames (below 40 dB).

The switched predictor coefficients are determined in accordance to the estimated correlation coefficient of the available input speech. Thus, for a_1 we adopted the value of average correlation coefficient calculated over all weakly correlated frames. Vice-versa, a_2 is taken to be the average correlation coefficient calculated over all highly correlated frames. For the tested speech signal, assuming $M = 80$ samples for the frame, we get $a_1 = 0.23$ and $a_2 = 0.95$.

Table I summarizes the average values of *SNR* obtained according to (24) for the proposed multilevel ΔM, for various frame lengths (i.e. $M = 80$, 160, 240 and 320 samples) and different number of quantization levels (32, 64, 128 or 256). One can observe that the highest *SNR* values in all considered scenarios is obtained for $M = 80$, which is expected as the quantizer codebook is adjusted more often. As a baseline, we also provide in Table I the results for the non-predictive case, i.e. forward adaptive scalar compandor (Fig. 1), denoted as SNR_{PCM}. The proposed multilevel ΔM is superior compared to the baseline, with 10 dB gain in *SNR*.

TABLE I. THE PERFORMANCE OF THE PROPOSED MULTILEVEL ΔM, FOR VARIOUS FRAME LENGTHS AND NUMBER OF QUANTIZATION LEVELS.

	$N = 32$			$N = 64$			$N = 128$			$N = 256$		
M	$SNR_{\Delta M}$	SNR_{PCM}	$R_{\Delta M}$	$SNR_{\Delta M}$	SNR_{PCM}	$R_{\Delta M}$	$SNR_{\Delta M}$	SNR_{PCM}	$R_{\Delta M}$	$SNR_{\Delta M}$	SNR_{PCM}	$R_{\Delta M}$
80	29.00	19.19	5.075	35.14	25.28	6.075	41.08	31.32	7.075	46.97	37.40	8.075
160	28.98	19.12	5.037	35.06	25.21	6.037	41.00	31.27	7.037	46.96	37.35	8.037
240	29.03	19.08	5.025	35.02	25.19	6.025	40.96	31.31	7.025	46.96	37.38	8.025
320	28.78	19.00	5.019	34.83	25.07	6.019	40.75	31.14	7.019	46.72	37.17	8.019

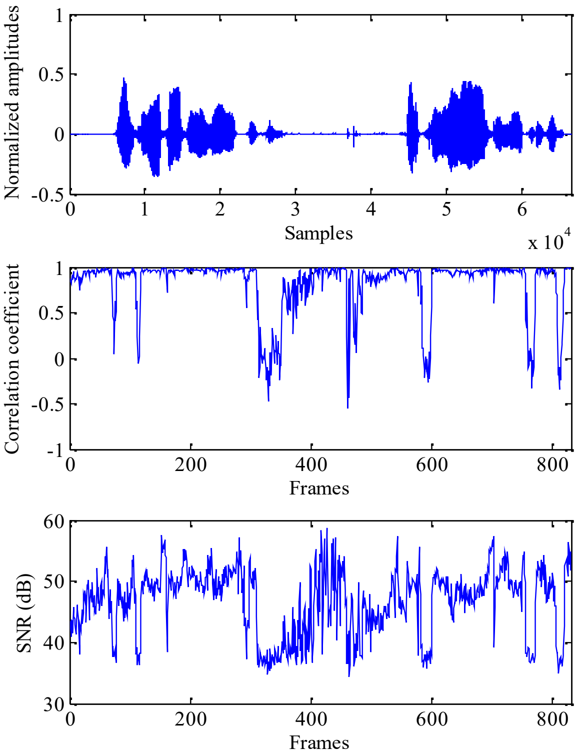


Fig. 6. The correlation coefficient ρ and *SNR* over all frames.

Note that the obtained experimental results are in agreement with the theoretical ones shown in Fig. 3–Fig. 5, indicating that there is a valid reason to apply the proposed solution in high-quality quantization of speech signal.

The complexity of the proposed algorithm remains unchanged compared to the baseline, it is equal to O(N^2).

V. CONCLUSIONS

In this paper, the speech coding algorithm based on the multilevel delta modulation and the first-order switched prediction based on correlation is considered. The obtained results indicate that the proposed algorithm offers significantly better signal quality compared to the G.711 algorithm, with about 10 dB gain in *SNR*. Moreover, since the proposed solution has a small complexity, it can be successfully employed for high-quality speech coding.

REFERENCES

[1] N. S. Jayant, P. Noll, *Digital Coding of Waveforms, Principles and application to Speech and Video*. New Jersey: Prentice Hall, 1984.

[2] W. C. Chu, *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. New Jersey: John Wiley & Sons, 2003.

[3] L. Hanzo, C. Somerville, J. Woodard, *Voice and Audio Compression for Wireless Communications*. London: John Wiley & Sons, 2007.

[4] ITU-T, Recommendation G.711. Pulse Code Modulation of Voice Frequencies, 1972.

[5] Y. Hiwasaki, H. Ohmuro, T. Mori, S. Kurihara, A. Kataoka, “A G.711 embedded wideband speech coding for VoIP conferences”, *IEICE Trans. Information and Systems*, vol. E89-D, no. 9, pp. 2542–2552, 2006. DOI: 10.1093/ietisy/e89–d.9.2542.

[6] Y. Hiwasaki, T. Mori, S. Sasaki, H. Ohmuro, “A wideband speech and audio coding candidate for ITU-T G.711 WBE standardization”, in *Proc. IEEE ICASSP*, Las Vegas, USA, 2008, pp. 4017–4020.

[7] Z. Peric, J. Nikolic, “An adaptive waveform coding algorithm and its application in speech coding”, *Digital Signal Processing,* vol. 22, no. 1, pp. 199–209, 2012. DOI: 10.1016/j.dsp.2011.09.001.

[8] Y. Hiwasaki, T. Mori, S. Sasaki, H. Ohmuro, A. Kataoka, “A wideband speech and audio coding candidate for ITU-T G.711 WBE standardization”, in *Proc. of the IEEE ICASSP,* 2008, pp. 4017–4020. DOI: 10.1109/ICASSP.2008.4518535.

[9] G. M. Petkovic, Z. Peric, L. Stoimenov, “Switched scalar optimal µ-law quantization with adaptation performed to both the variance and the distribution of speech signal”, *Elektronika ir Elektrotechnika*, vol. 22, no. 1, pp. 64–67, 2016. DOI: 10.5755/j01.eee.22.1.14111.

[10] V. Despotovic, Z. Peric, L. Velimirovic, V. Delic, “DPCM with forward gain-adaptive quantizer and simple switched predictor for high quality speech signals”, *Advances in Electrical and Computer Engineering*, vol. 10, no. 4, pp. 95–98, 2010. DOI: 10.4316/aece.2010.04015

[11] Z. Peric, B. Denic, V. Despotovic, “Delta modulation system with a limited error propagation”, in *Proc. XIII Int. Conf. SAUM*, Nis, Serbia, 2016.

[12] J. D. Gibson, “Speech compression”, *Information*, vol. 7, no. 32,

pp. 1–22, 2016. DOI: 10.3390 /info7020032.

[13] V. Prochazka, P. Pollak, J. Zdansky, J. Nouza, "Performance of Czech speech recognition with language models created from public resources", *Radioengineering,* vol. 20, pp. 1002–1008, 2011.

[14] D. G. Zrilic, *Circuits and Systems Based on Delta Modulation*. New York: Springer, 2005, ch. 1.

[15] A. Ortega, M. Vetterly, "Adaptive scalar quantization without side information", *IEEE Trans. Image Processing*, vol. 6, no. 5, pp. 665–676, 1997. DOI: 10.1109/83.568924.

[16] D. Aleksic, Z. Peric, J. Nikolic, "Support region determination of the quasilogarithmic quantizer for Laplacian source", *Przeglad Elektrotevhniczny*, vol. 88, no. 7, pp. 130–132, 2012.

[17] R. Goldberg, L. Riek, *A Practical Handbook of Speech Coders*. CRC Press, 2000.