www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2020, VOL. 66, NO. 1, PP. 69-76
Manuscript received November 15, 2019; revised January, 2020. DOI: 10.24425/ijet.2019.130267

Low Delay Sparse and Mixed Excitation CELP Coders for Wideband Speech Coding

Przemysław Dymarski

***Abstract*—Code Excited Linear Prediction (CELP) algorithms are proposed for compression of speech in 8 kHz band at switched or variable bit rate and algorithmic delay not exceeding 2 msec. Two structures of Low-Delay CELP coders are analyzed: Low-delay sparse excitation and mixed excitation CELP. Sparse excitation is based on MP-MLQ and multilayer models. Mixed excitation CELP algorithm stems from the narrowband G.728 standard. As opposed to G.728 LD-CELP coder, mixed excitation codebook consists of pseudorandom vectors and sequences obtained with Long-Term Prediction (LTP). Variable rate coding consists in maximizing vector dimension while keeping the required speech quality. Good speech quality (MOS=3.9 according to PESQ algorithm) is obtained at average bit rate 33.5 kbit/sec.**.

***Keywords*—CELP, Low-Delay CELP, MP-MLQ, MOS, variable bit rate**

I. Introduction

In comparison of speech coders the following issues are considered: signal quality, bit rate, algorithmic delay, computational complexity. Narrowband speech (bandwidth less than 4 kHz) is nowadays judged as low quality, therefore wideband speech (bandwidth 7-8 kHz) [1,2,3,4,5,6,7,8] or even full-band speech (14-22 kHz) [9,10] is processed in telecommunication services. In real time services, like VoIP, delay is an important issue. Most speech coding algorithms introduce delay of some tens of milliseconds, but multiple of this delay is observed due to transmission, buffering and decoding. In some applications, like networked music performance or audio-conference echo delay should not exceed 30 ms [11].Thus algorithmic delay of a coder should be limited to some milliseconds.

Unfortunately most of standard wideband speech and audio coders, operating at sampling frequency 16 kHz, exhibit substantial algorithmic delays (Table I). Narrowband variant of G.711.1 coder has one-sample delay, but the wideband variants (bit rates 80 and 96 kbit/s) have about 12 ms delay. Among the wideband G.722 coders only the simplest one, two band ADPCM coder, exhibits low delay (4 ms), but at high bit rate (64 kbits/s). Newer variants operate at lower bit rates but their algorithmic delay exceeds 25 ms.

High delay of some wideband coders stems from transform coding (MDCT) which requires large block of samples for transform calculation and quantization. This concerns G.718 [6], G.729.1 [5] and EVS [2] coders.

Przemysław Dymarski is with Warsaw University of Technology, Institute of Telecommunications, Poland (e-mail: dymarski@tele.pw.edu.pl).

Besides of G.722 coder [7] low delay of 5 ms shows the BroadVoice coder designed for VoIP applications [1]. It uses a specific CELP algorithm, called two-stage noise feedback coding (TSNFC). Due to vector quantization, 80 samples of excitation signal are encoded in only 120 bits. Predictive filter, pitch and gains require only 40 bits, thus 160 bits per 80 samples yields bit rate of 32 kbit/s.

TABLE I
COMPARISON OF LOW DELAY WIDEBAND CODERS

coder	delay [ms]	bit rate kbit/s
G.711.1	12	80, 96
G.722	4	64
G.722.1	40	24 - 32
G.722.2	25	16
G.718	42	8 - 32
G.729.1	49	14 - 32
EVS	20	6.6 - 24
BroadVoice	5	32
Opus (WB)	26.5	16 - 64

Opus coder [10] consists of two algorithms: SILK (a specific CELP coder) and CELT (MDCT transform coder). Good quality of speech and music is assured at delay of 26.5 ms, but there is a variant of SILK with delay of 5 ms and CELT with delay of 8.7 ms. It should be noted that the low delay CELT is a full-band coder operating at 44100 samples per second, so delay is equal to 256 samples [9].

Target of this paper is to propose algorithms for wideband speech coding at low delay of 1-2 ms. For VoIP applications scalable coders are required, able to switch between several bit rates, due to varying quality of transmission channels [12]. In packet transmission variable rate coder may be applied, where each packet refers to speech signal frame of different duration. For some kind of audioconferences embedded coders are required, in which low rate bitstreams are hidden in high rate bitstream. Thus participants of an audioconference may use transmission channels of different throughput.

Due to low delay and scalability requirements transform coding algorithms are not considered in this paper. Delay of a typical CELP coder (Fig.1a) also exceeds 20 ms. Speech is processed using codebook vectors containing about 5 ms of excitation signal, but much longer frame is required to calculate prediction coefficients, describing the synthesis filter $H(z)$. Due to pitch predictor called also a long-term predictor (LTP) excitation signal of the synthesis filter becomes quasi-

periodic which is favorable for encoding of voiced speech. An example of such coder is G.722.2 wideband CELP coder of delay equal to 25 ms [3].

In order to decrease delay, backward adaptation of synthesis filter H(z) is applied. In consists in using decoded signal x^* instead of original speech signal x for calculation of prediction coefficients. Such predictor describes past signal frames and it follows changes of speech signal with substantial delay. However, backward adaptation introduces no algorithmic delay. Delay depends only on dimension of codebook vectors. In a narrowband CELP coder G.728 vector dimension is 5 and algorithmic delay is reduced to 0.625 ms [13]. To prevent from increase of bit rate there is no pitch prediction in G.728 standard coder (Fig.1b). The proposed low delay algorithms for wideband speech coding are based on the structure of G.728 coder with substantial modifications. Dimension of processed vectors is equal to N=16, which yields algorithmic delay of 1 ms. Scalability is obtained with application of K-sparse excitation (K nonzero components in N-dimensional vector of excitation signal) and variable K (Fig.1c). For calculation of sparse excitation Multipulse Maximum Likelihood Quantization (MP-MLQ) algorithms were used [14,15]. Optimal sparse excitation was tested using a modified Sphere Decoding algorithm [16]. Finally multilayer sparse excitation was synthesized, based on the ideas expressed in [12] and [15]. The proposed algorithm is scalable and it may be applied also in variable rate and embedded coders.

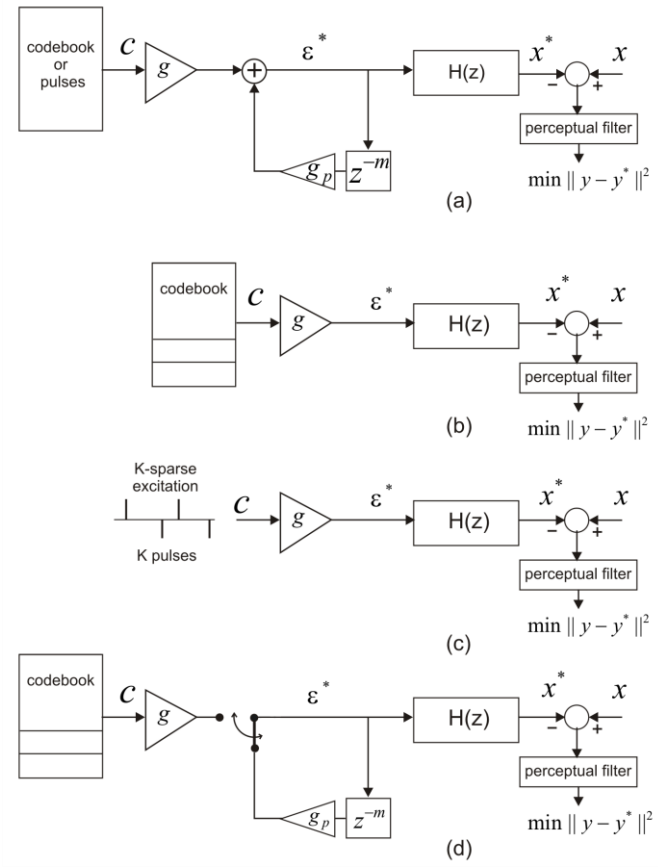

Fig. 1. Typical CELP coder (a), Low-delay CELP according to G.728 standard (b), proposed Low-delay sparse excitation CELP coder (c) and Low-delay mixed excitation CELP (d)

Scalability may also be obtained by changing dimension of codebook vectors. In the extreme case, codebook contains scalar values and the algorithm is equivalent to ADPCM. Codebook usually contains N-dimensional vectors uniformly distributed on a surface of N-dimensional sphere. For simplification of calculation of excitation vector c only one vector is issued from the codebook, like in G.728 coder. Improvement of voiced speech coding may be obtained by application of pitch predictor. However, using pitch predictor according to Fig.1a would practically double the bit rate. Therefore, coder structure presented in Fig.1d is proposed. Excitation signal is taken from a codebook or is searched in the past with a pitch predictor. Speech coder using pitch predictor only was called a self-excited vocoder [17] and idea of mixing different kinds of excitation signals was first expressed in [18]. In this paper these approaches are tested in a wideband low delay coder. Finally a variable rate coder is obtained by maximizing vector dimension while keeping the required speech quality. It is shown that variable rate coder yields better speech quality than the constant rate coders at the same bit rate.

This paper is organized as follows: In Section II low delay sparse excitation CELP coder is described (Fig.1c) and tested. In Section III mixed excitation low delay CELP is presented (Fig.1d) and compared with the sparse excitation CELP. Variable rate coder, based on the mixed excitation CELP algorithm, is described in Section IV. Final conclusions are presented in Section V.

II. Low Delay CELP Coders with Sparse Excitation

CELP coder is based on the analysis-by-synthesis approach. Many vectors of excitation signal $\varepsilon*$ are tested so as to minimize distance between the original speech vector x and decoded speech vector $x*$ appearing at the output of predictive synthesis filter:

$$H(z)=\frac{1}{A(z)}=\frac{1}{1-P(z)}=\frac{1}{1-\sum_{m=1}^{p}a_m z^{-m}} \qquad (1)$$

Spectral weighting of quantization noise is attained by using the perceptual filter (here of transfer function $A(z)/A(z/\gamma)$, $\gamma \simeq 0.95$), enabling greater distortion in formant regions, according to masking threshold. Excitation vectors are selected so as to minimize the squared Euclidean distance between vectors of perceptual signals: $\|e\|^2=\|y-y*\|^2$. At successive stages of modeling the spectral flatness of the error signal e increases and the quantization noise accompanying the output speech signal $x*$ attains its proper spectral shape.

In order to reduce delay, backward adaptation of predictive synthesis filter $H(z)$ is applied, like in G.728 narrowband standard coder. Linear prediction coefficients are calculated using decoded speech $x*$ multiplied by a window shown in Fig.2. Duration of the window is equal to 20 ms, which is typical in speech processing, but the fact, that delayed and quantized speech is used for predictor calculation is not favorable for speech quality. On the other hand, backward predictor adaptation is used with ADPCM coders.

Exponentially decreasing window is typically used in sequential adaptation algorithms implemented in these coders, but in Low-delay CELP better speech quality was obtained using window shown in Fig.2. Main advantage of backward adaptation, besides of low delay, is lack of transmission of prediction coefficients. Therefore, more bits may be destined to encode excitation signal.

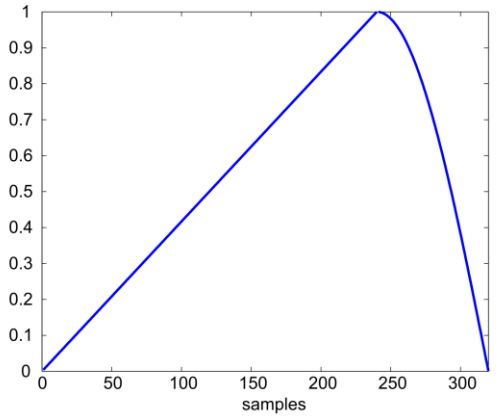

Fig. 2. Window applied for backward adaptation of predictive synthesis filter H(z)

In the proposed Low-delay CELP coder the excitation signal is calculated in vectors of dimension N=16, yielding algorithmic delay of 1 ms. K-sparse excitation is applied, $K<N$ components of excitation vector have nonzero values. Several excitation models are used in CELP coders:

a) Multipulse Excitation (MPE):

$$\varepsilon^* = \sum_{i=1}^{K} g_i \, c^{j(i)}, \qquad K < N \tag{2}$$

where c^j - pulse of unit amplitude at position j, $1 \le j \le N$,
$j(i)$ - position of pulse number i
g_i - gain (amplitude) of pulse number i

b) MP-MLQ and ACELP excitation, applied in narrowband CELP coders, e.g. in GSM-EFR, GSM-AMR, G.729, G.723.1 standards [19]:

$$\varepsilon^* = g \sum_{i=1}^{K} s_i \, c^{j(i)} \tag{3}$$

where $s_i = \pm 1$ - polarity (sign) of the pulse
g - common gain for all pulses.

There are two variants of this scheme. In the MP-MLQ *(Multi-Pulse - Maximum Likelihood Quantizer*) there are no restrictions or small restrictions concerning positions of the selected vectors (pulses), e.g. in the G.723.1 coder operating at bit rate 6.3 kbit/s either even or odd positions may be taken. In the ACELP (*Algebraic CELP*) coders pulses are distributed in tracks and have usually 8-16 possible positions within a vector of dimension N=40-60. Such excitation is used e.g. in the G.723.1 coder operating at 5.3 kbit/s [19].

c) multilayer , used in the G.718 coder [6], [12]:

$$\varepsilon^* = \sum_{l=1}^{L'} g_l \sum_{i=K'(l-1)+1}^{K'l} s_i \, c^{j(i)} \tag{4}$$

Here, there are L' layers in which K' pulses are distributed as in (3). In each layer, however, a separate gain g_l is used.

Pulses positions and amplitudes should now be found, so as to minimize the distance between perceptual vectors $\| e \|^2 = \| y - y^* \|^2$. In excitation model (3) for a given gain g there are $\frac{N!}{(N-K)!\,K!} 2^K$ possible excitation signals (for N=16 and K=8 it is about 3.3 millions). Testing of all possible combinations of pulse positions and signs is not feasible, so many suboptimal algorithms are proposed. In [15] these algorithms are implemented in a narrowband high delay CELP coder and compared. Greedy algorithms consist of K iterations and yield one pulse per iteration. Such algorithms are very simple, but excitation vector computed in this way is far from being optimal.

The M-best approach consists in allocating, in a parallel way, M sequences of pulses. At the first step (k=1) the excitation signal consisting of one pulse is considered. N pulse positions are sorted in ascending order according to the approximation error $\| e \|^2$. The first M vectors start M sequences. At the k^{th} step there are almost MN possible sequences (to any of M sequences any of N-k+1 pulses may be appended), but only M of them are retained. Permutations of the same pulse positions are eliminated. At the last step only one sequence is selected.

Pulse positions and signs may be then recalculated using replacement algorithms. Each pulse, one by one, is replaced to its better position, if such position exists. The criterion is minimum of approximation error. This procedure is repeated in a cyclic manner. If in K trials there is no effective replacement (each pulse stays at its previous position) then the algorithm is stopped.

In the proposed Low-delay wideband coder, both approaches are combined: M-best and replacement. Predictive synthesis filter and perceptual filter is described with 20 prediction coefficients, calculated synchronously at coder and decoder side using backward adaptation algorithm. Number of pulses (K nonzero components in a N=16-dimensional vector) equals 2,4,6,8 or 10. Thus different bit rates may be obtained (Fig.3). MP-MLQ excitation model was used (3) and gain was encoded in $b_g = 4$ bits on logarithmic scale. Thus number of bits required for coding one frame (vector) of signal is equal to

$$B_{MLQ} = \left\lceil \log_2 \left(\frac{N!}{(N-K)!K!} \right) \right\rceil + K + b_g \tag{5}$$

In Fig.3 (continuous line) segmental signal to quantization noise ratio is drawn for a phrase of Korean speech. SNR_{seg} is the average value of signal power $\| x \|^2$ to quantization error power $\| x - x^* \|^2$ ratio calculated in segments (16-dimensional vectors) and expressed in decibels. Saturation of SNR_{seg} is observed if number of pulses exceeds K=8. Indeed, number of possible pulse configurations decreases and for greater values of K number of bits per frame (5) drops and so does SNR_{seg}.

Some improvement is obtained using optimal excitation signal (Fig.3, dashed line). Due to very high complexity (over 3 million of searches for K=8 and N=16) this is not a real time

algorithm. It may be made faster if Sphere Decoding (SD) algorithm is applied. Sphere Decoding consists in testing only these solutions which yield approximation error $\|e\|^2$ less than the best solution known so far. Some modification of SD algorithm is required, in order to calculate not only the positions and signs of K pulses but also the gain g. Such modification is described in [16]. As in full search algorithm, the optimal positions and signs of K pulses are obtained, but complexity is considerably reduced: at K=8 and N=16 mean value of searches is about 20000. Some disadvantage of this algorithm is its variable complexity. In some cases the optimal solution is found in several hundreds of searches, but in some rare cases number of searches may attain a million. However, search may be interrupted in these cases yielding the best solution obtained so far.

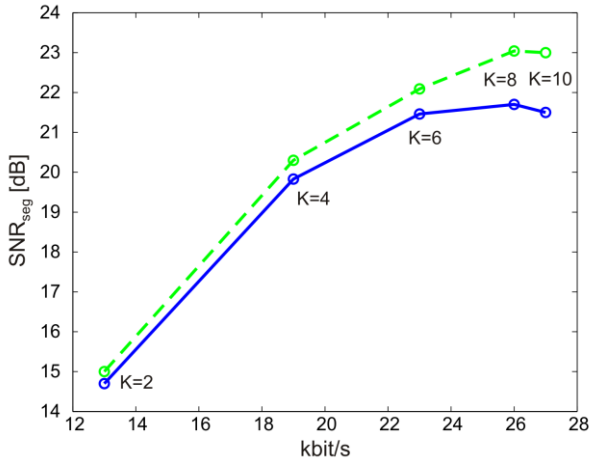

Fig. 3. Segmental SNR for a phrase of Korean speech: blue continuous line - M-best with replacement, dashed green line – modified Sphere Decoding

Results presented in Fig.3 show that MP-MLQ excitation model (3) does not assure required tradeoff: better signal quality at greater bit rate. Solution of this problem is a multilayer excitation model proposed in [12]. Multilayer excitation (4) is a mixture of MPE (2) and MP-MLQ (3) signals.

In the proposed wideband Low-delay coder different sparse excitation model is applied: up to 10 pulses MP-MLQ model is used, then each pulse obtains its proper gain, like in MPE model:

$$\varepsilon^* = g\sum_{i=1}^{K} s_i\, c^{j(i)} \qquad if\ K \le 10$$
$$\varepsilon^* = g\sum_{i=1}^{10} s_i\, c^{j(i)} + \sum_{i=11}^{K} g_i\, s_i\, c^{j(i)} \qquad if\ K > 10 \tag{6}$$

Gains for MPE are encoded in 3 bits each. Tests of the proposed sparse excitation was performed using 12 phrases of male and female speech in Polish, English, Korean, German, Danish and Italian. Mean value of segmental SNR is presented in Fig.4 and MOS obtained with PESQ algorithm [20] in Fig.5. Now the required tradeoff is achieved: better speech quality is obtained at a cost of greater bit rate.

The proposed sparse excitation model may be implemented in a scalable CELP coder, in which the required bit rate is obtained using appropriate number of pulses (K). Bit rate may be also changed at every frame (16 samples). Variable rate CELP obtained in this manner will have greater bit rate, because value of K should be transmitted in every frame. After some modifications, e.g. using greedy algorithm for pulses allocation, the proposed algorithm may be also implemented in embedded CELP coder. To each decoder different number of pulses could be transmitted in this case.

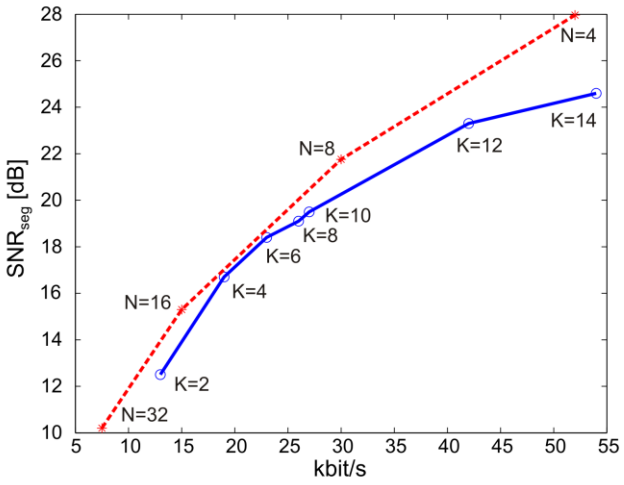

Fig. 4. Segmental SNR for 12 phrases of speech: blue continuous line – proposed sparse excitation, dashed red line – mixed excitation

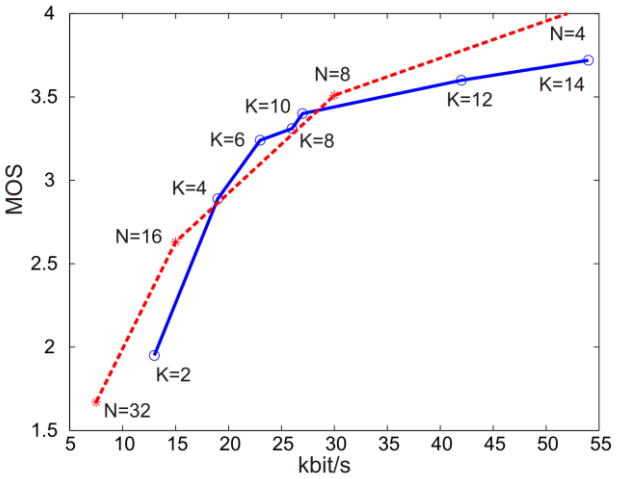

Fig. 5. MOS values for 12 phrases of speech: blue continuous line – proposed sparse excitation, dashed red line – mixed excitation

III. Low Delay CELP Coders with Mixed Excitation

Scalability of LD-CELP coder may be also obtained using variable dimension of processed vectors. In this case a wideband counterpart of a narrowband G.728 standard is obtained (Fig.1b). In order to simplify the codebook search algorithm, only one vector is selected from a codebook and multiplied by the gain ($g > 0$). Algorithmic delay of the proposed wideband coder is equal to dimension of codebook vector N. Bit rate depends on dimension of these vectors, number of vectors in the codebook (L) and number of bits for gain coding (b).

$$R = \left(\lceil \log_2 L \rceil + b\right)\frac{16000}{N} \tag{7}$$

Parameters of simulated Low-delay coders are shown in Table II. If two codebooks are used, then the number of vectors in (7) is $L=L_1+L_2$.

TABLE II
PARAMETERS OF TESTED LD-MIX CODERS

N	L_1	L_2	b	R kbit/s	SNR [dB]	MOS
1	0	2	4	80	32.2	4.06
2	0	16	4	64	28.8	4.01
4	0	256	5	52	27.9	4.00
8	256	256	6	30	21.8	3.51
16	256	256	6	15	15.3	2.63
32	256	256	6	7.5	10.2	1.67

If vector dimension is reduced to 1, then Low-delay CELP becomes ADPCM coder. Indeed, in the linear predictive filter (1) the sample of excitation signal ε^* is added to the output of predictor $P(z)$ and thus the output speech sample x^* is produced. In the same way the quantized prediction error is added to the predicted sample to obtain the output sample of ADPCM decoder. Algorithm of excitation signal calculation plays role of a quantizer.

Codebooks of the proposed Low-Delay coders consist of L_2 normalized vectors, uniformly distributed on the surface of the N-dimensional sphere. In case of ADPCM coder there are 2 scalar values, +1 and -1. For higher vector dimension the following codebook design algorithm is applied:

- Generation of 100 L_2 pseudorandom sequences of Gaussian pdf,
- Normalization of these N-dimensional vectors,
- Clustering using K-means algorithm to obtain L_2 centroids
- Normalization of centroids which become codebook vectors.

Only one vector c is selected from the codebook, then it is multiplied by the gain $g>0$. Gain is quantized using 2^b quantization levels. Because of high dispersion of gain values, logarithmic quantizer was used (linear on decibels scale). Predictive gain coding may lead to reduction of number of bits b shown in Table II by one.

In a typical CELP coder (Fig.1a) Long Term Predictor (LTP) is used, in order to improve coding of voiced speech. LTP uses correlation of a current speech vector with a speech vector delayed by pitch period or its multiple. In Fig.1a prediction of the current speech vector x is obtained with delayed excitation signal multiplied by LTP gain g_p and filtered by $H(z)$. Several delays (m) are tested. Delayed excerpts of excitation signal ε^* form a series of vectors, belonging to so called adaptive codebook. Application of LTP requires encoding of two parameters: delay m and gain g_p. In a Low-delay CELP coder based on G.728 standard, where only index of selected vector (j) and its gain (g) are transmitted, it would double the bit rate. So as to avoid this, new structure of LD-CELP is proposed (Fig.1d). Excitation vector is searched in two codebooks: constant one and adaptive one. Only one vector is chosen, which minimizes approximation error $\|e\|^2=\|y-y^*\|^2$. Such mixed excitation CELP coder (LD-MIX CELP) requires only one more bit to encode N-dimensional vector of speech if number of vectors in constant and adaptive codebooks are the same ($L_2=L_1$).

LTP was applied for vector dimension $N\geq 8$ (Table II). For encoding of LTP gain g_p the same number of bits (b) is used as for encoding of gain g. However, quantization levels are not the same. Due to quasi-periodic character of voiced speech, values of g_p are close to 1. In transient segments different values of LTP gain appear, so uniform linear quantizer is used with quantization levels from 0 to 3.

SNR_{seg} and MOS values are obtained using a long speech file being a concatenation of 12 speech phrases mentioned in Section 2. They are presented in Table II and in Fig.4 and Fig.5. The mixed-excitation LD-CELP outperforms sparse excitation LD-CELP. Using a typical high delay CELP coder (Fig.1a) better speech quality is obtained at the same bit rate. In [21] wideband LD-CELP of delay equal to 0.5 ms was simulated and compared with a typical CELP of delay equal to 16 ms [22]. Similar MOS values were obtained for Low delay CELP at 30 kbit/s and high delay CELP at 24 kbit/s.

What is the impact of LTP on speech quality in wideband LD-CELP? To answer this question, let us observe which codebook, constant or adaptive, is used more frequently. For most processed speech phrases, the adaptive codebook was used 2 times more frequently than the constant one. In Fig.6 indexes of selected LTP vectors are shown for a phrase of Korean speech (delay m is equal to index plus N). In most cases, delays are equal to pitch period and its multiples. The other experiment consisted in comparing LD-CELP with two codebooks ($L_1=L_2=256$ vectors) with LD-CELP with only a constant codebook, but containing $L=512$ vectors, thus having the same bit rate. Vector dimension was equal to $N=16$, 12 speech phrases were concatenated to obtain a speech file. With LTP segmental SNR value was improved by 1.15 dB and MOS by 0.3.

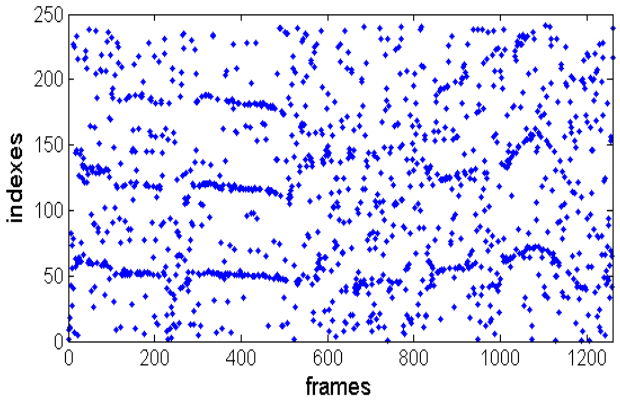

Fig. 6. Indexes of vectors selected from adaptive codebook

Using only LTP without constant codebook is called a self-excited vocoder (SEV) [17]. The idea of mixing different kinds of excitation signals was expressed in [18]. However, these approaches were not used in Low-delay CELP coders.

IV. VARIABLE RATE LOW DELAY CELP CODERS

MOS values presented in Table II and Fig.5 suggest, that vectors of dimension 16 and 32 do not assure acceptable

quality of speech signal. On the other hand, SNR values calculated for every 32-dimensional vector (Fig.7) indicate quite good quality (SNR > 20 dB) for many excerpts of speech signal. The same phrase was encoded using 8-dimensional vectors (Fig.8). For some speech excerpts SNR attains 40-50 dB, so bit rate could be reduced without loss of perceived speech quality. These observations suggest application of variable bit rate coding.

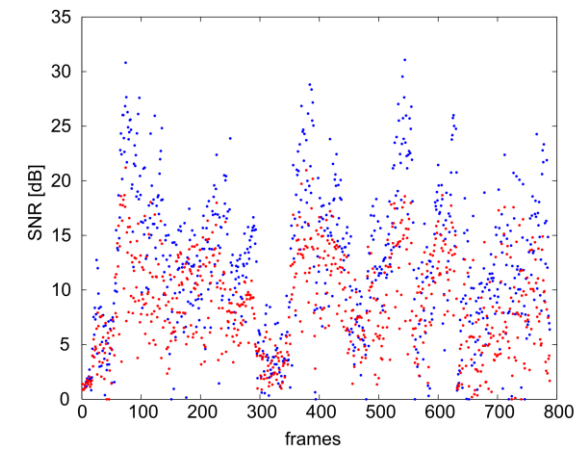

Fig. 7. SNR calculated for 32-dimensional vectors of speech signal (blue) and perceptual signal (red) – Korean speech phrase

Firstly all the coders mentioned in Table II were included in the proposed variable rate low-delay wideband speech coder, but finally ADPCM coder (N=1) and CELP coder processing N=2-dimensional vectors were rejected, because signal quality obtained at N=4 was sufficient in most cases. Dimension is allocated to each frame of speech, so as to minimize bit rate and maintain acceptable speech quality. As a quality measure signal to quantization noise ratio at perceptual signal level was used (Fig.1):

$$SNR_{perc} = \frac{\| y \|^2}{\| y - y^* \|^2} \tag{8}$$

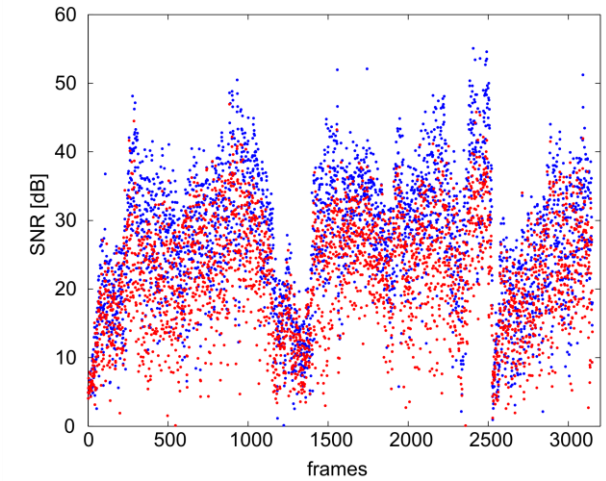

Fig. 8. SNR calculated for 8-dimensional vectors of speech signal (blue) and perceptual signal (red) – Korean speech phrase

At the first stage of vector dimension allocation algorithm maximum dimension N=32 is tested, in order to minimize bit rate. This defines algorithmic delay of the whole algorithm, because 32 speech samples should be buffered, even though lower dimension is finally selected. 32 samples correspond to delay of 2 ms. The best vector is searched in both codebooks containing 32-dimensional vectors. For the best vector, minimizing $\| e \|^2 = \| y - y^* \|^2$, SNR_{32} is calculated (8). If $SNR_{32} > T_{32}$, dimension N=32 is accepted, 32 samples of speech are encoded and appropriate packet is transmitted. If SNR_{32} is too low, then vectors of dimension 16 are processed. In the same way, the best vector is searched in both codebooks. If $SNR_{16} > T_{16}$, then dimension N=16 is accepted. If not, then the best 8-dimensional vector is searched. If $SNR_8 < T_8$, then CELP coder processing 4-dimensional vectors is used.

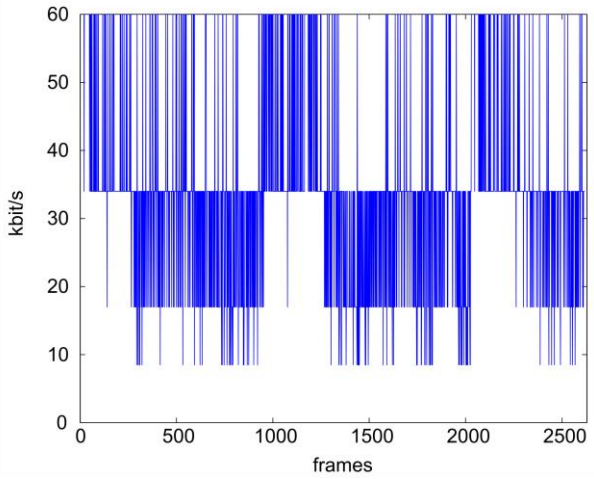

Fig .9. Bit rate variations in variable bit rate coding of Korean speech phrase

Each processed vector may have different dimension: 32, 16, 8 or 4. Side information of 2 bits should be appended to each transmitted packet. Thus bit rate corresponding to four available dimensions equals 8.5, 17, 34 and 60 kbit/s (compare with values in Table II). Bit rate varies rapidly in time, see Fig.9.

Performance of the proposed variable rate coder depends on thresholds T_{32}, T_{16} and T_8. Lower threshold values yield lower bit rate and worse quality of speech. SNR values for vectors of variable dimension are shown in Fig.10 for thresholds T_{32}=T_{16}=20 dB and T_8=10 dB. Note that segments of low quality (SNR < 10 dB) are rare, compare with Fig.7 and Fig.8.

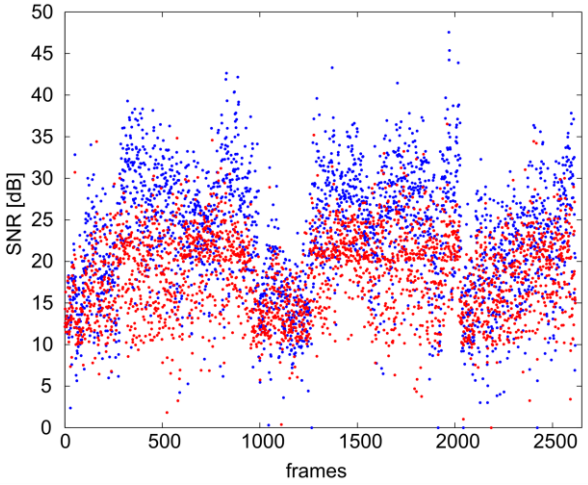

Fig. 10 SNR calculated for vectors of variable dimension (blue – speech signal, red - perceptual signal) – variable rate coding of Korean speech phrase

In Fig.11 MOS values for 7 wideband speech phrases are compared for constant and variable rate coding. MOS was evaluated using PESQ standard algorithm [20]. Thresholds in variable rate coding were T_{32}=T_{16}=20 dB and T_8=10 dB. Despite of side information yielding an increase of bit rate, variable bit rate coder outperforms constant bit rate coders.

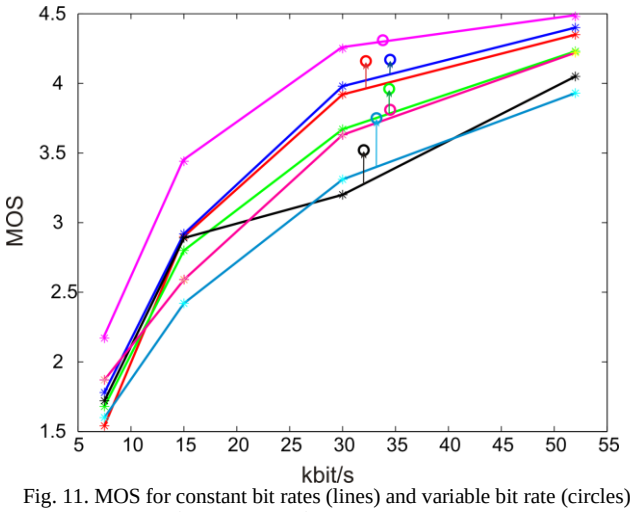

Fig. 11. MOS for constant bit rates (lines) and variable bit rate (circles) for 7 phrases of wideband speech

MOS values vary with speech phrase, for variable bit rate coder from 3.5 to 4.3 at similar average bit rate. Using the speech file obtained by concatenation of 12 speech phrases spoken in different languages MOS=3.9 was obtained at average bit rate 33.5 kbit/s.

V. Conclusions

Wideband speech coding problem at low delay is analyzed in this paper. There are not many algorithms of this kind, inter alia, the BroadVoice coder, having algorithmic delay of 5 ms at bit rate 32 kbit/s [1]. In this article two kinds of CELP algorithms are described, having different excitation signal of predictive synthesis filter: sparse excitation and mixed excitation.

New form of sparse excitation is proposed, based on popular MP-MLQ algorithm combined with Multipulse Excitation (MPE). Thus a scalable coder is obtained, operating at many bit rates and offering a tradeoff: better speech quality at greater bit rate (Fig.4 and Fig.5). Algorithmic delay of this coder is equal to 1 ms. The proposed sparse excitation CELP coder may be implemented as a variable rate coder. Moreover, it has an emdedded structure: speech of lower quality may be decoded using only a part of excitation signal. Some disadvantage is its computational complexity - about 200 Mflops. However, it is not a problem for signal processing technology nowadays.

The proposed mixed excitation CELP is based on a narrowband G.728 standard coder. In its new structure (Fig.1d) two kinds of excitation signals are switched: signal from adaptive codebook and non-adaptive codebook. It is proved that this kind of excitation performs better at the same bit rate than signal from non-adaptive codebook only. Bit rate depends on dimension of processed vectors, using different dimension a scalable coder is obtained, yielding better speech quality than the sparse excitation coder (Fig.4 and Fig.5). This coder was also simulated as a variable rate coder of delay equal to 2 ms. Despite of side information necessary for transmitting varying dimension of processed vectors, variable rate coder yields better speech quality than constant bit rate coder at similar average bit rate (Fig.11). MOS value (obtained using PESQ standard algorithm [20]) for speech file being a concatenation of 12 phrases spoken in 6 languages, was equal to 3.9 at the average bit rate 33.5 kbit/s.

Tests of wideband BroadVoice coder, described in [1], yielded average MOS=3.79 at bit rate 32 kbit/s. MOS was also calculated with PESQ algorithm, but speech database was much wider.

Advantage of the proposed variable rate low delay speech coder is its low computational complexity, several tens of Mflops. Some disadvantage is lack of embedded structure – it is not possible to decode speech having only a part of a bitstream.

References

[1] Chen Juin-Hwey and J. Thyssen, “The Broadvoice Speech Coding Algorithm*”. IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP2007*, pp.537-540, DOI 10.1109/ICASSP.2007.366968

[2] ETSI. “3GPP TS 26.441 EVS codec”, 2014.

[3] ITU-T, “Recommendation G.722.2, Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)”, 2003.

[4] ITU-T, “Recommendation G.722.1, Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss”, 2005.

[5] ITU-T, “Recommendation G.729.1:G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729”, 2006

[6] ITU-T, “Recommendation G.718, Frame error robust narrow-band and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s”, 2008.

[7] ITU-T, “Recommendation G.722, 7 kHz audio-coding within 64 kbit/s”, 2012.

[8] ITU-T, “Recommendation G.711.1: Wideband embedded extension for ITU-T G.711 pulse code modulation”, 2012.

[9] J.M. Valin, T.B. Terriberry, C. Montgomery and G. Maxwell, “A High-Quality Speech and Audio Codec With Less Than 10 ms Delay”. *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 1, Jan. 2010, DOI 10.1109/TASL.2009.2023186

[10] K. Vos, K. V. Sorensen, S. S. Jensen and J.M. Valin “Voice coding with Opus” *135th AES Convention*. 2013

[11] Z.Kurtisi; X. Gu and L. Wolf, "Enabling network-centric music performance in wide-area networks". *Communications of the ACM*. 49 (11) 2006, pp.52–54, DOI 10.1145/1167838.1167862

[12] J.Stachurski, “Embedded CELP with adaptive codebooks in enhancement layers and multi-layer gain optimization”, *Proc. ICASSP 2009*, pp.4133-4136, DOI 10.1109/ICASSP.2009.4960538

[13] ITU-T, “Recommendation G.728, Coding of speech at 16 kbit/s using low-delay code excited linear prediction”, 2012.

[14] F. K. Chen, G. M. Chen, B. K. Su and Y. R. Tsai, “Unified pulse replacement search algorithms for algebra codebooks of speech code”, *IET Signal Proc.*, 2010, Vol. 4, Iss. 6, pp. 658-665, DOI 10.1049/iet-spr.2009.0216

[15] P.Dymarski, R.Romaniuk "Sparse Signal Modeling in a Scalable CELP Coder", *Proc.21st European Signal Processing Conf. EUSIPCO 2013*, Marrakech, Morocco, We-P.1.1, ISBN 978-1-4799-3687-8

[16] P.Dymarski, R.Romaniuk, "Modified Sphere Decoding Algorithms and their applications to some sparse approximation problems", *Proc. 22nd European Signal Processing Conf. EUSIPCO 2014*, Lisbon, DOI 10.5281/zenodo.43826

[17] R. Rose and T. Barnwell “The self-excited vocoder - an alternate approach to toll quality at 4800 bps”. *IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '86.*

[18] P. Dymarski and N. Moreau. "Mixed excitation CELP Coder". *Proc. European Conference on Speech Communication and Technology (EUROSPEECH'89)*, Paris 1989

[19] ITU-T, "Recommendation G.723.1, Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s", 2006.

[20] ITU-T, „Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs", 2001.

[21] K. Kim, "Wideband LD-CELP coder" – *BS thesis WEiTI, Warsaw University of Technology*, supervisor P. Dymarski, 2019

[22] G. Kim, "Wideband speech coding using CELP algorithm" – *BS thesis WEiTI, Warsaw University of Technology*, supervisor P.Dymarski, 2019