Perceptually Constrained Variable Bitrate Wideband Speech Coder

Michael Livshitz, Alexander Petrovsky *

Abstract — **A high quality wideband speech coder based on code-excited linear prediction (CELP) algorithm with perceptually constrained variable bitrate (VBR) is proposed in this paper. A VBR is achieved with the help of reconfigurable structure of multiband multistage codebook of excitation vectors controlled by psychoacoustic model based on Warped Discrete-Fourier Transform (WDFT). Comparison with MPEG1 Layer III codecs at 16, 24 and 32 kbps is implemented.**

Keywords — **CELP, Multiband Codebook with Multistage Structure, Psychoacoustics, Reconfigurable Codebook Structure, Variable Bitrate, WDFT.**

I. Introduction

NOWADAYS much attention is paid to wideband (7kHz) speech coders development. It is connected with such bandwidth expansion advantages as better quality of reconstructed speech, intelligibility and capability of transmitting some properties of the acoustic environment and music in this frequency range. Wideband CELP coders are traditionally divided into two classes: split-band [1] and fullband [2]. In split-band CELP coding, which belongs to the class of subband coders, each subband signal is encoded by a CELP coder. The split-band CELP coders generally suffer from degradation of speech quality in the frequency ranges where the responses of the filterbanks overlap. In contrast to split-band CELP coding, a fullband coding scheme directly encodes wideband signals using the only CELP coder. The fullband CELP coders usually suffer from a high frequency noise in the decoded speech. The majority of modern coders have so-called layered structure [3]. It means that each layer is responsible for coding in the given frequency range (subband) and quantity of layers is determined by a bandwidth of the signal being encoded. Such scheme has certain drawbacks, because it lacks flexibility and does not reflect features of human auditory system. In other words, it has a perceptual redundancy. Applying a psychoacoustic model for control structure of such coders presents one of the novel trends in the sphere. As a result, a bit allocation strategy for coded layers is received. The actual bit allocation for each band depends then on the target bitrate and its ordering. For example, with 8 bands, there are 40320 possible ordering combinations that are never implemented. At the same time, there is a number of ordering combinations that appear frequently. So, we have to make bit reallocation in bit reservoir according to available strategies and in order of perceptual significance.

The coder presented in this paper combines high compression ratio of linear prediction with perceptual coding capabilities of a transform coder. To eliminate abovementioned drawbacks of described coders, we propose a wideband coder with hybrid structure, which integrates the main principles of the listed approaches with the modern concepts in psychoacoustic theory. The coder is based on multiband excitation CELP algorithm with multistage quantization of excitation signal under perceptually monitored structure of the multiband multistage codebook. Psychoacoustic model is based on Warped Discrete-Fourier Transform (WDFT) [4] approximating a bark scale [5] which is used for subband decomposition to get subband codebooks training sets. To avoid the problem of bit reallocation discussed earlier, vector quantization of codebook structure is employed. As recent researches show, this approach allows very effective and precise quantization of codebook structure using 10 bits only. In contrast to conventional wideband coders at fixed bitrates, the proposed model provides better quality of reconstructed speech at significantly decreased bitrate. It demonstrates the quality comparable with MPEG1 Layer III [6] codec at 24 kbps.

II. Operation of the Coder

A. Structure of the coder and principles of multistage vector quantization of excitation signal

The structure of proposed coder is shown in Fig. 1. Input signal is quantized by 16 bit per sample with sampling frequency f_S=16kHz. LP analysis is implemented for every frame after preemphasis of the input signal. Inverse and weighting filters are cascaded and used to obtain weighted residual signal S_W, which is divided into 4 subframes of 5 ms length each to estimate a long-term predictor (LTP) parameters. The order of short-term predictor (STP) is p=16.

A core of the coder is a multiband codebook of excitation vectors with multistage structure. The codebook is formed in so-called “off-line” mode from bandpass-filtered vectors. Subband decomposition is implemented by a non-uniform cosine modulated polyphase filterbank [7] according to Table 1 and is used to prepare the subband training sets.

* The authors are with Computer Engineer Department, Belarusian State University of Informatics and Radioelectronics, 6, P. Brovka st., Minsk, Belarus (phone: +375 17 231-29-10; fax: +375 17 239-84-20; e-mail: mlivshitz@tut.by; palex@bsuir.by)

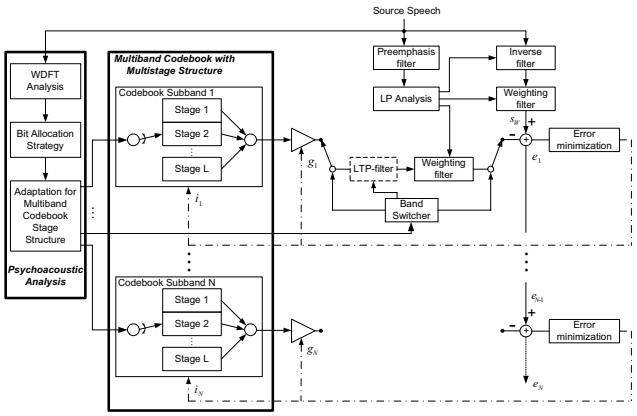

Fig. 1. Block diagram of proposed wideband CELP coder.

TABLE I: SUBBAND DECOMPOSITION.

Subband	Freq. range, Hz	Barks
1	100-510	4
2	510-1080	4
3	1080-1720	3
4	1720-2320	2
5	2320-3150	2
6	3150-4100	1.5
7	4100-5300	1.5
8	5300-8000	3

The psychoacoustic model balanced with accepted critical frequency scale of subband decomposition and structure of multistage vector quantization of subband excitation signal [8] is embedded into the coder to eliminate perceptual redundancy of coded signal. The unit of psychoacoustic analysis determines an optimal structure of the codebook for considered frame that is based on estimation of subband perceptual entropy (SPE). SPE is transformed to search depth in subband codebooks (Fig. 1) by algorithm described in [9]. WDFT analysis buffer contains previous history and currently encoded frame of original speech and is used to avoid rapid changes of codebook structure. The buffer has length of 512 samples: initial 192 samples contain the last part of the previously encoded frame and the remaining samples contain currently encoded frame of the original signal. Only subbands determined by the unit of psychoacoustic analysis are involved into the coding process. If two first subbands are among the coding subbands than the LTP-filter is involved into encoding to get a fine-tuned formant structure of speech. The procedure adds 2.6 kbps to a total bitrate (see Table 2, strings in italic). Search of subband excitation vectors is implemented sequentially in order of perceptual importance of subbands, from low frequency to high frequency bands (Fig. 1). Cascaded LTP and weighting filter are connected to the current stage of vector quantization of excitation signal. The cycle of minimization makes an error minimization loop in which the contribution of the current subband is subtracted from weighted residual signal of the previous stage e_{i-1} to form signal e_i that will be quantized at the next stage of the multistage quantization scheme. During multistage vector quantization a kind of coded signal "whitening" is made, i.e. with each stage of quantization the flatness of its spectrum is increased, and, therefore, statistical redundancy decreases: the signal becomes noise-like and its quantization needs smaller quantity of bits. Algorithm is repeated for all subframes of the currently processed frame. It is necessary to underline that filtered code vectors have final length (which equals subframe length of 5 ms). It leads to a "leakage" of spectral energy between adjacent subband codebooks. However, as code vectors of different codebooks are nearly-orthogonal, sequential search in each codebook provides almost the same quality as optimal joint search in all subband codebooks, but with significant complexity reduction. Moreover, the "leakage" of spectral energy of *i*-th subband can be compensated at quantization of *i+1*-th subband, therefore there is no need for a paraunitary filterbank.

B. Parameters quantization

LP coefficients are transformed to line spectral frequencies (LSF) and the split vector is quantized by splitting on three subvectors of dimension 3, 3, 10, accordingly. Each of the subvectors is quantized with 9 bits. The minimized norm of error vector takes into account the distance between LSFs, SPE value and the decrease in frequency resolution of human ear for high frequencies.

In general, the structure of multiband codebook can be different to meet the requirements of coding bandwidth, throughput capacity of communication channel and quality of the reconstructed speech. The following modification of the codebooks is accepted for proposed coder: multistage subband codebook with 5-stage organization and dimension for each stage of 16, 32, 64, 128, 256 vectors (4, 5, 6, 7, 8 bit-per-stage) accordingly. Structure of the codebook is vector quantized with 10-bits codebook, which was trained on a 20-minute database material of speech and music signals. The scheme of parameters coding is shown in Table 2.

TABLE II: PARAMETERS QUANTIZATION SCHEME.

Model parameter	Param./ frame	Bits/ param.	Bits/ frame	Frames/s	Bitrate, bps
LSF	16	1.6875	27	50	1350
Model Gain	1	7	7		350
LTP Delays	*4*	*8*	*32*		*1600*
LTP Gains	*4*	*5*	*20*		*1000*
Codebook Structure	1	1.25	10		500
Excitation Gains	0 – 32	4	0 – 128		**0 – 6400**
Book Indexes	0 – 32	4 – 8	0 – 256		**0 – 12800**
Peak bitrate, bps					**24000**

Codebook structure, indexes, excitation gains, STP and LTP parameters are transmitted in the decoder (Fig. 2), where they are restored. Subband excitation signals are formed, summed and shaped by synthesis filter. Adaptive post-filtering and deemphasis of reconstructed speech signal are performed at the final stage.

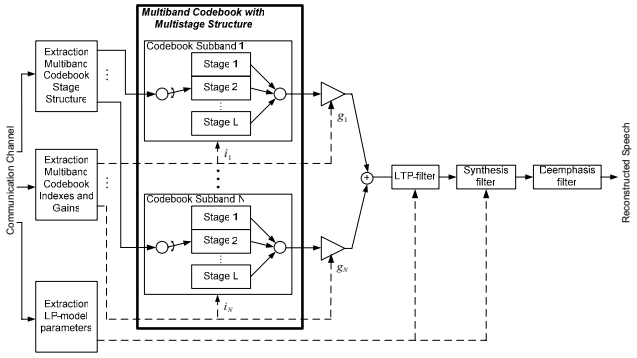

Fig. 2. Structure of decoder.

As it is shown in Table 2, the bitrate is changed according to the characteristic of coded frame of signal and can achieve peak value of 24 kbps for the codebook structure with fully employed subbands.

Using the coder with limited (fixed) structure of the codebook eliminates the necessity of storing the configuration information for multiband codebook with multistage structure, and, consequently, the maximum bitrate will reach 23.5 kbps.

Let us consider configuration of multiband codebook with multistage structure for two adjacent frames of speech signal (Fig. 3, Fig. 4). Fig. 3 and Fig. 4 show that codebook structure precisely reflects all features of the signal coded frame.

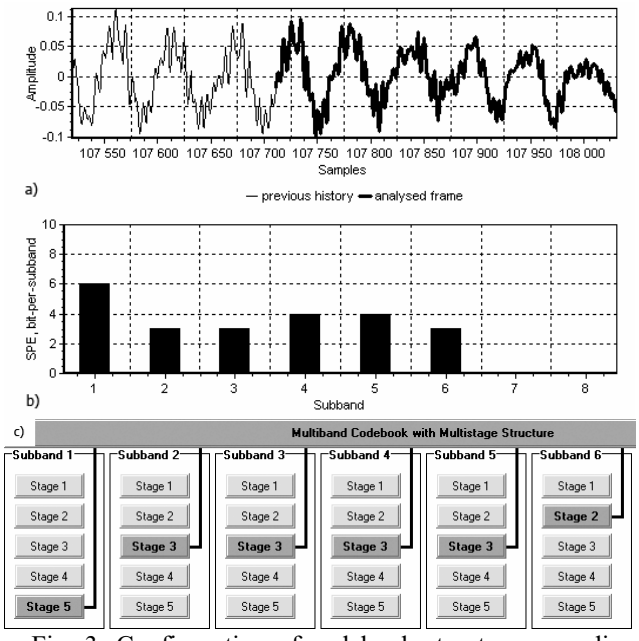

Fig. 3. Configuration of codebook structure according to the perceptual redundancy of the i-th speech frame (a) buffer of the analysis, (b) SPE value, (c) codebook structure.

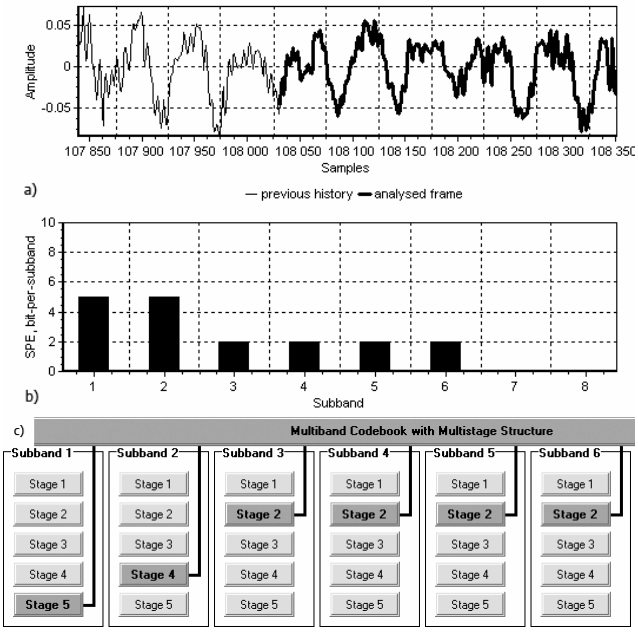

Fig. 4. Configuration of codebook structure according to the perceptual redundancy of the i+1-th speech frame (a) buffer of the analysis, (b) SPE value, (c) codebook structure.

Subband perceptual entropy estimation in each of the subbands determines the number of subbands used in the coding process, and also defines search depth (number of stages involved) in subband codebooks.

III. QUALITY EVALUATION

For an estimation of quality of the proposed coder a test material from the TIMIT [10] database was used. The total duration of the material is 5 minutes. Estimation technique described in [11]-[14] was applied. For comparison purposes MPEG1 Layer III [6] codecs at 16, 24, 32 kbps and the proposed coder with fully subband employed and reconfigurable codebook structure were used.

Noise-to-Mask Ratio (*NMR*) [12], Signal-to-Noise Ratio (*SNR*), Bark Spectrum Distortion (*BSD*) [13] and Modified Bark Spectrum Distortion (*MBSD*) [14] were estimated. Finally, the MOS values were evaluated. The results are shown in Table 3.

As an example, in Fig. 5 the spectrograms of an original and reconstructed signal of MPEG1 Layer III at 24 kbps and the proposed coder with variable bitrate are shown. The test material consists of a phrase (first 3.5 seconds) and a music fragment.

As illustrated in Fig. 5, bitrate of the proposed coder varies according to a perceptual significance of the coded signal (Fig. 5b). The analysis of spectrograms (Fig. 5 c, d, e) proves the preservation of spectral components present in the original signal.

The MOS evaluations of compared coders are presented in Fig. 6. The proposed coder at 23.5 kbps provides quality which is comparable with MPEG1 Layer III codec at 32 kbps, and VBR version has almost the quality of MPEG1 Layer III at 24 kbps.

TABLE III: OBJECTIVE SPEECH QUALITY EVALUATION.

Codec	NMR_{total}, dB	NMR_{seg}, dB	SNR_{total}, dB	SNR_{seg}, dB	*BSD*	*MBSD*
MPEG1 Layer III at 16 kbps	1.69	0.55	1.53	1.37	1.02	0.09
MPEG1 Layer III at 24 kbps	1.48	0.19	1.94	1.56	0.99	0.08
MPEG1 Layer III at 32 kbps	1.43	0.11	2.27	2.28	0.96	0.08
Proposed coder at 23.5 kbps	-6.95	-7.01	7.62	6.07	0.97	0.06
Proposed VBR coder	-6.15	-5.78	5.63	3.97	1.00	0.05

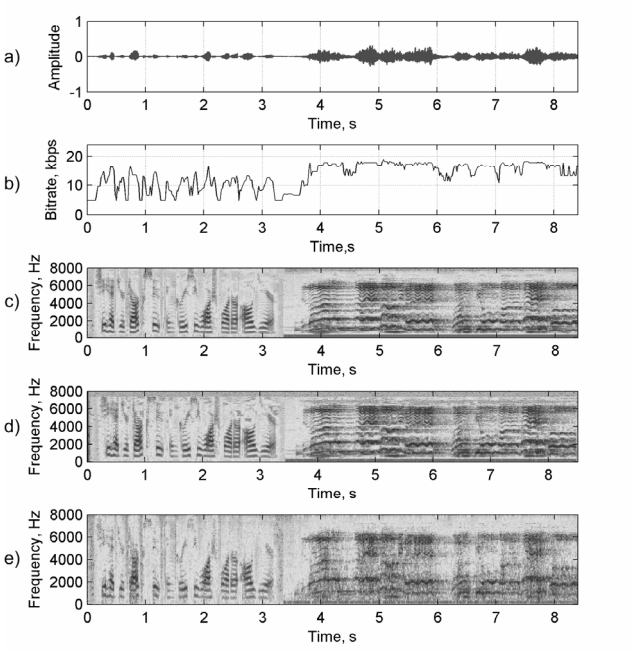

Fig. 5. Comparison of quality of the reconstructed signal (a) original signal waveform, (b) bitrate variation in the proposed coder, (c) spectrogram of the original signal, (d) spectrogram of the reconstructed signal (MPEG1 Layer III at 24 kbps), (e) spectrogram of the reconstructed signal (proposed VBR coder).

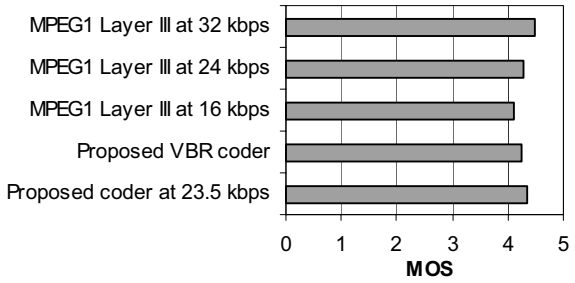

Fig. 6. Comparison of codecs' subjective quality.

IV. CONCLUSION

Experimental results testify the high quality of the proposed wideband VBR CELP coder. Estimations from Table 3 demonstrate that proposed VBR coder provides quality comparable with MPEG1 Layer III codecs, and on *NMR* and *SNR* parameters significantly outperforms them. The major advantages of the proposed coder are reconfigurable structure of the codebook and, as a result, variable bitrate and computational complexity; high compression ratio of LP with perceptual coding capabilities of a transform coder. The preliminary estimation of computational complexity has shown that proposed coder in full-duplex mode can effectively be realized on the base of the Texas Instruments TMS320C6x processor with VLIW-architecture. Thus, the rest of the processor resources can be utilized for pre- and post-processing systems to further improve of the speech quality.

REFERENCES

[1] R. Drago, R. Montagna, F. Perosino, and D. Sereno, "Some experiments of 7-kHz audio coding at 16 kbits/s," in P*roc. of the IEEE International conference on Acoustic, Speech, Signal processing, ICASSP*, pp.192-195, 1989.

[2] A. Ubale and A. Gersho, "A multi-band CELP wideband speech coder," in P*roc. of the IEEE International conference on Acoustic, Speech, Signal processing, ICASSP*, pp.1367-1370, 1997.

[3] C. Erdmann and P. Vary, "Embedded Speech Coding on Pyramid CELP," in Proc. of IEEE Workshop on Speech Coding, Tsukuba, Ibaraki, Japan, October 2002, pp.29-31.

[4] M. Parfieniuk, A. Petrovsky, "Warped DFT as the basis for psychoacoustic model," in P*roc. of the IEEE International conference on Acoustic, Speech, Signal processing, ICASSP*, vol. IV, Montreal, Canada, May 2004, pp.185-188.

[5] E. Zwicker, H. Fastl, "Psychoacoustics Facts and Models," Springer-Verlag, Berlin Heidelberg, 1990.

[6] ISO/IEC JTC1/SC29/WG11, MPEG, International Standard IS 13818-3 Information technology – Generic Coding of Moving Pictures and Associated Audio, 1994.

[7] A. Petrovsky, M. Parfieniuk, K. Bielawski, "Psychoacoustically Motivated Non-uniform Cosine Modulated Polyphase Filter Bank," in Proc. of the 2nd International Workshop on Spectral Methods and Multirate Signal Processing (SMMSP 2002), Toulouse, France, September 7-8, 2002, pp.95-101.

[8] M. Z. Livshitz, M. Parfieniuk, A. A. Petrovsky, "Multistage Vector Quantization with Variable Dimension in Perceptual Speech Encoders with Psychoacoustic Model Based on Warped DFT," in Proc. of the 7th International Conference on Digital Signal Processing and its Applications, vol.VII-1, Moscow, Russia, 2005, pp.187-191.

[9] M. Livshitz, M.Parfieniuk, A. Petrovsky, "Wideband CELP-encoder with Multiband Excitation and Multistage Quantization under Reconfigurable Structure Codebook," *Digital Signal Processing*, Moscow, Russia, to be published.

[10] TIMIT Acoustic-phonetic continuous speech corpus, NISTIR 4930.

[11] Al. A. Petrovsky, "Objective Speech Quality Evaluation of Reconstructed Audio Signal of Perceptual WP-Encoder Based on the Peripheral Model of Human Ear," in Proc. of the 5th International Conference on Digital Signal Processing and its Applications, vol. V-2, Moscow, Russia, 2003, pp.123-126.

[12] K. Brandenburg, T. Sporer, " "NMR" and "Masking Flag": Evaluation of Quality Using Perceptual Criteria," in Proc. of the 11th Int. Conv. Aud. Eng. Soc., "Test and measurement", Portland, USA, May 1992, pp.169-179.

[13] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," IEEE J. Select. Areas Commun., vol. 10, pp. 819-829, June 1992.

[14] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of a modified bark spectral distortion measure as an objective speech quality measure," IEEE ICASSP, pp.541-544, Seattle, 1998.