International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:2, No:7, 2008

High Quality Speech Coding using Combined Parametric and Perceptual Modules

M. Kulesza, G. Szwoch, and A. Czyżewski

***Abstract*—**A novel approach to speech coding using the hybrid architecture is presented. Advantages of parametric and perceptual coding methods are utilized together in order to create a speech coding algorithm assuring better signal quality than in traditional CELP parametric codec. Two approaches are discussed. One is based on selection of voiced signal components that are encoded using parametric algorithm, unvoiced components that are encoded perceptually and transients that remain unencoded. The second approach uses perceptual encoding of the residual signal in CELP codec. The algorithm applied for precise transient selection is described. Signal quality achieved using the proposed hybrid codec is compared to quality of some standard speech codecs.

***Keywords*—**CELP residual coding; hybrid codec architecture; perceptual speech coding; speech codecs comparison.

I. Introduction

TRANSMISSION of speech signal through the Internet is used in a wide range of modern applications. In spite of constant development of Internet technologies, transmission of high bit-rate speech signals in PCM format is not practical, because it results in high delays in communication. This is especially important in bi-directional communication, e.g. in VoIP (Voice over IP) telephony, because high bit-rate means that the user has to wait longer for the response from the other speaker. This results in low quality of the service assessed by its users. Therefore, speech compression algorithms have to be applied in order to reduce bit-rate and decrease delays.

In most of the current applications related to transmission of speech signal, parametric coding algorithms are used (CELP, ACELP, LD-CELP, etc.). These algorithms reduce bit-rate of the signal significantly, sacrificing quality of the signal to some degree. For many years, bit-rate and delay were the main criteria in speech codec assessment, while subjective signal quality, expressed using the mean opinion score (MOS) scale, was considered less important. Most of parametric speech codecs used in current applications provides signal quality from 3.2 to 4.0 in MOS scale (where 5.0 means best possible quality) [1].

In terms of constant development of computer technologies, it seems that bit-rate and delay criteria in speech codecs assessment will be gradually becoming less important, and signal quality will become a decisive factor for comparison of various speech coding algorithms. Thus, it seems reasonable to consider a novel speech codec architecture that will sacrifice low bit-rate to some extent in order to improve signal quality significantly.

The experiments presented in this paper were focused on modification of existing parametric speech coding algorithms in order to improve subjective signal quality. Although the intelligibility of speech signal encoded by CELP codecs is usually satisfactory, the highest quality possible to obtain in that codec architecture is still limited. Instead of various techniques utilized during analyze-by-synthesis procedure, the one straightforward method of improving the codecs quality is additional encoding of the encoder residual signal [2]. This kind of codec architecture has been previously proposed as an efficient method of lossless wideband speech coding [3]. Although the lossless speech coding is essential for storage and future editing in the recording or in movie industry, it is not crucial for telecommunication. Hence, the focus of this research is to investigate the lossy speech codec by providing the enhancement layer for an existing standardized CELP codecs. In order to improve the overall quality of speech signal the residual signal is encoded with the usage of the perceptual module. While proposed architecture is similar to the one utilized in MPEG-4 CELP scalable speech codec [4], the method of the residual signal coding employed in our experiments is different.

The main problem in the parametric approach to speech coding is how to encode transients, voiced and unvoiced signal components, efficiently. Encoding of transient states is especially important here, because inappropriate encoding of transients results in significantly decreased signal quality, Various parametric codecs use different approach to this problem, yet none of these approaches provide sufficiently accurate transient encoding, which is reflected in quality values (MOS). One of the concepts of hybrid codec presented in this paper is extraction of transient, voiced and unvoiced components from the signal and using an appropriate approach for each of these groups. Certain literature references may be found on using the "sine-transient-noise" model for the wide-band audio signals (mostly for music) [5]. In the synthesis of musical instruments, introduction of transient analysis and synthesis to the "sine and noise" model resulted in improved signal quality. It may be expected that using a similar "voiced-transient-unvoiced" approach to speech signal will provide an improvement of signal quality, as well. However, no research on this topic has been done so far.

The aim of the hybrid "parametric-perceptual" speech

codec is to improve signal quality by incorporating perceptual coding algorithm into the parametric codec. Various speech coding algorithms use different approach to the problem of reducing signal bit-rate. For the purpose of experiments described in this paper, CELP parametric codec was selected as a base for development of the hybrid codec. The CELP architecture, with some modifications, is used in most of modern speech coding applications.

II. Voiced/Unvoiced Parts Selection Algorithm

Since the LP coding relies on a simple two-state model of speech production, each frame of input signal is classified either as a voiced or unvoiced. Usually, the classification is based on the observation that frames of voiced parts are strongly correlated with each other and have relatively higher energy than unvoiced parts [6]. This approach is also utilized in the engineered algorithm. However, the additional logic module is employed in order to ensure that frames classified as unvoiced do not contain transients. The diagram illustrating the engineered algorithm for selecting voiced and unvoiced parts of the speech signal is presented in Fig. 1.

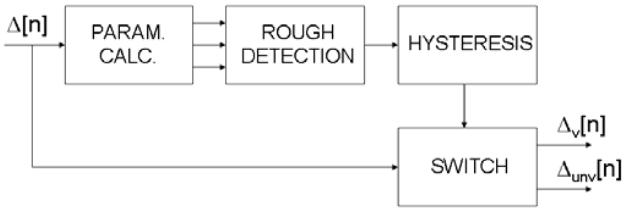

Fig. 1 Block diagram of the voiced/unvoiced selector

The "rough detector" relies on three parameters which are calculated for every block of segmented signal according to the following formulas [6][7]:

$$x_o = \frac{1}{2N}\sum_{n=1}^{N-1}\left|\mathrm{sgn}\left(s[n]\right)-\mathrm{sgn}\left(s[n-1]\right)\right| \quad (1)$$

$$x_1 = \frac{1}{N}\sum_{n=0}^{N-1}\left|s[n]\right| \quad (2)$$

$$x_2 = \max\left(\sum_{n=0}^{\frac{N}{2}-1} s[n]\cdot s\left[n+\frac{N}{2}\right]\right) \quad (3)$$

where: $s[n]$ – block of the signal, N – frame length

The frame is classified as voiced if the following expression is true:

$$w_0 + \sum_{k=1}^{M} w_k \cdot x_{k-1} > 0 \quad (4)$$

where: w_k – elements of weighting vector, M – number of parameters

The w_k elements of weighting vector were chosen in order to allow a proper frames classification for different speech samples. Since the rough detector does not take into account any information about the previous classification results, the voiced/unvoiced decision may change instantly from one frame to another when some special conditions occur. Thus, an appropriate hysteresis function is utilized in order to prevent undesirable state changes of the detector. It has to be mentioned that not only pure-voiced frames but also frames containing transients are classified into the voiced part of the speech signal. The hysteresis module controls the state of the switch in order to define the current frame as a voiced or unvoiced, and also to trigger the appropriate fade-in and fade-out operation if the transition of the detector state occurs. The result of classification for a particular speech sample is presented in Fig. 2.

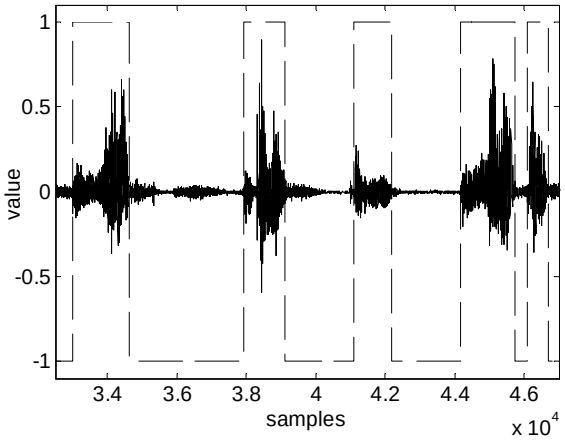

Fig. 2 Results of voiced/unvoiced classification

An important insight is that involving the triangular window into the cross-mixing procedure induces phase distortions into the signal. Thus, the window based on the cosine function was utilized. The fade-in and fade-out functions are defined as follows:

$$winFin = \frac{1}{2}\left(1-\cos\left(\frac{n\pi}{N-1}\right)\right) \quad (5)$$

$$winFout = \frac{1}{2}\left(1-\cos\left(\frac{(N-1-n)\pi}{N-1}\right)\right) \quad (6)$$

where: n=0, 1, …, N-1; N – frame length

The selected voiced and unvoiced parts of the particular speech sample are shown in Fig. 3. It can be noticed from Fig. 2 and Fig. 3 that the engineered algorithm is effective in selecting the voiced and unvoiced parts of the speech signal.

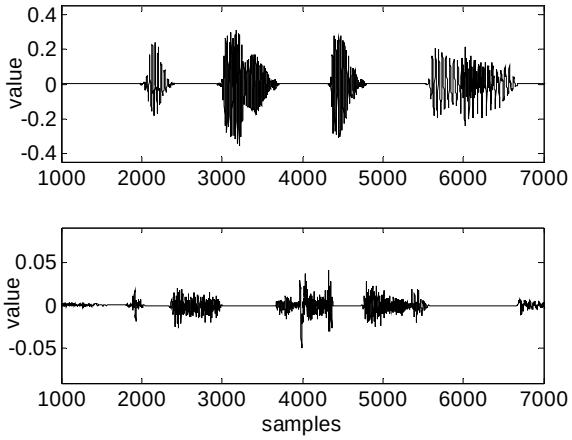

Fig. 3 Voiced (top) and pure unvoiced (bottom) parts of speech signal

III. Transient Selector Algorithm

The proposed algorithm consists of two main stages. In the first one, the rough voicing decision is made, and in the second one, both pure-voiced part and transients of the signal are selected. The traditional approach to transient detection is based on the assumption that the energy of the signal increases rapidly when the transient occurs [8]. Although the energy tracking of the signal is useful for detecting the time-domain transients, it fails in case of frequency-domain transients occurrence. If the frequency-domain transient occurs, the energy distribution changes over frequency, while total energy of the signal remains nearly constant. In order to detect that group of transients, it is necessary to analyze the energy variations in some subbands instead of tracking the energy of entire signal [9]. The diagram illustrating the engineered algorithm for time-domain and frequency-domain transients selection is shown in Fig. 4. The $s_{tr}[n]$ and $s_{pv}[n]$ signals represent the transient and pure-voiced part of the input signal respectively.

The transient selector operates as follows. In the first stage, the input signal is divided into some short segments, and for each segment the Fourier spectrum is calculated (with the use of the FFT algorithm). The part of the spectrum representing frequencies above 100 Hz is then divided into N uniform subbands, and for each subband the energy en(b) is obtained. It has been found empirically that analysis of energy variations in eight subbands is sufficient for a transient detection in speech signals. In the next step, the value of f(b) for each subband is calculated. The f(b) function is formulated in the following way,

$$f_n(b) = \alpha \cdot en(b) + (1-\alpha) \cdot f_{n-1}(b) \quad (7)$$

where: α – constant

Further, the parameters related to transient measures are obtained according to the formula:

$$G(b) = \frac{en(b)}{f_n(b)} \quad (8)$$

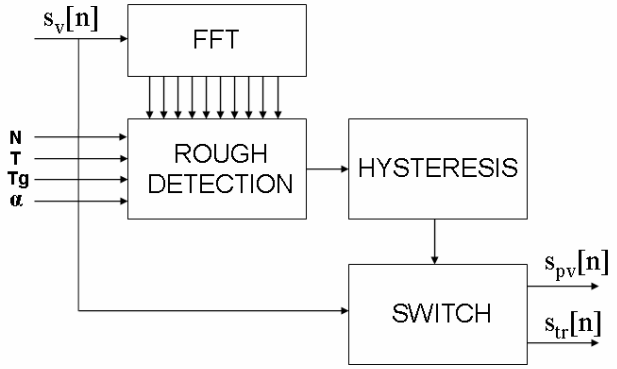

Fig. 4 Block diagram of transient selector

Next, the value of the G(b) parameter is compared with the T threshold for each subband and the total transient measure parameter F is calculated,

$$F = \sum_{b=0}^{N-1} d(b) \quad (9)$$

where: $d(b) = \begin{cases} 1 & G(b) > T \\ 0 & G(b) \le T \end{cases}$

In the last step, the value of F parameter is compared with the global threshold T_g. If $F>T_g$ the particular frame is classified as a one containing transient.

It has to be mentioned that the described algorithm is incorporated in the MPEG-4 AAC general audio encoder [10]. Since in this approach the band-limited speech signal is analyzed, additional parameterization is necessary in order to allow robust transients detection. An important observation is that the time-domain transients of speech signals are associated with transitions between unvoiced and voiced parts. Thus, the one way to make the detection algorithm more efficient is to measure the zero crossing rate in parallel [9]. Because that operation can be viewed as simple voicing detection, similar results may be yield when only the voiced part of the speech signal is fed into the transients detector [2]. Therefore in practical experiments first the voiced/unvoiced decision had been made, and then the transients were detected within the voiced part of the speech signal. The results of transient detection for a particular speech sample are depicted in Fig. 5.

It can be noted from the Fig. 5 that the proposed algorithm is able to assign signal segments containing both: time-domain and frequency-domain transients. However, due to limited resolution of analysis, some fragments of the transients may be placed in the neighborhood of the assigned frames. Therefore, an additional hysteresis module is employed in order to prevent transients segmentation during selection process. When the particular frame is determined as one

containing transient, the fade-in operation is applied to the previous one. The signal segments are then moved to the $s_{tr}[n]$ until the detector state changes. After that one more frame is moved and the fade-out operation is applied to the next frame. Although that procedure introduces an additional delay, it prevents transients segmentation efficiently.

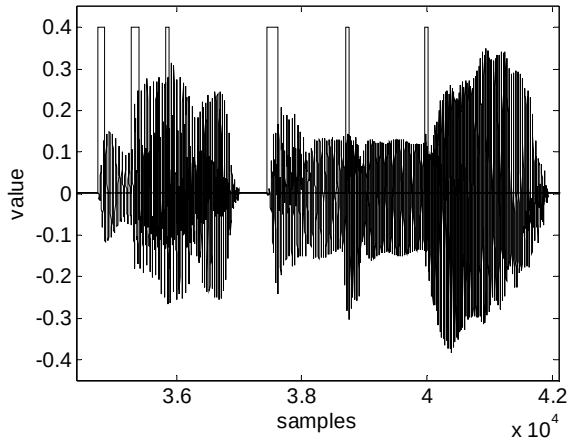

Fig. 5 Results of transient detection

IV. Perceptual Coding of the CELP Residual Signal

It is well known that in the analysis-by-synthesis loop the CELP codec is based on, the goal is to choose the appropriate excitation sequence and other parameters in order to match as much as possible the synthetic and original speech signals. The error between the synthesized and original signal is calculated for all excitation vectors in the codebook, and then the error-minimizing criterion is applied. It is worth mentioning that error signal is perceptually weighted in order to explore the masking phenomena of the human auditory system. The perceptual weighting filter attenuates the error signal spectrum in formant regions, while amplifying the error signal spectrum in non-formant regions [6]. Consequently, the residual signal of the entire CELP encoder adopts a formant structure and follows the shape of the input signal [3]. An analysis of the G723, G728 and G729 residual signals confirms those expectations.

It has to be mentioned, however, that perceptual coding of the residual signal with bit-rate significantly higher than bit-rate of the CELP module would be inefficient. Unfortunately, some substantial distortions are introduced to the noise-like parts of the residual signal during the perceptual coding with reduced bit-rate. In this case the overall quality can be even lower for the sum of degraded residual and input signal encoded by CELP than for the CELP signal itself. Therefore, in proposed approach only the voiced part of the residual signal is perceptually encoded. The architecture of the encoder proposed here is presented in Fig. 6. The input signal s[n] is first encoded by the CELP encoder operating at fixed bit-rate. The $s_{CELP}[n]$ signal represents the main bit stream, which is also decoded locally in order to obtain the residual signal $\Delta[n]$. In the next stage, the voiced and unvoiced parts of the residual signal are detected. It has to be noted that only the voiced part $\Delta_v[n]$ is further processed. As a sequel, the selected voiced part of residual signal is perceptually encoded and sent to the decoder in parallel to the CELP bit stream. The decoding process consists of two stages. In the first one, CELP and residual bit streams are decoded using the CELP and perceptual decoders, respectively. Next, the resulting signals are added together in order to compose the entire decoded speech signal.

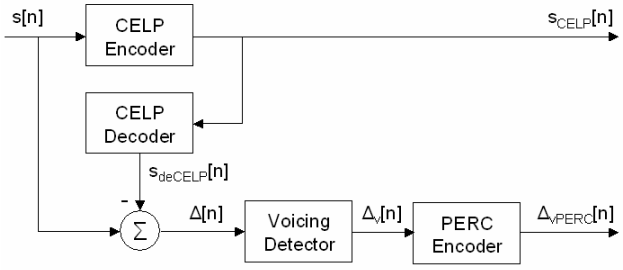

Fig. 6 Encoder architecture

During the experiments, the G728 LD-CELP codec operating at 12.8 kbps rate was utilized in a base layer of the codec. The residual signal was encoded employing the open source Ogg Vorbis perceptual module operating at 14 kbps average rate. Notwithstanding Ogg Vorbis is designed for compression of music and general audio, and is not optimized for perceptual speech coding, it seems to be useful in evaluating of the proposed codec concept [11].

The PESQ measurements were taken in order to asses quality gain achieved by the additional encoding of the voiced part of the residual signal. The partial PESQ scores were estimated for female and for male speech samples encoded with an usage of standardized G728 and proposed codec. In the next step partials PESQ scores for G728 and proposed codec were subtracted in order to calculate the partial PESQ gains. In Fig. 7, the waveform of the particular voiced part of residual signal together with a corresponding partial PESQ gains function are shown.

As it can be observed from the Fig. 7, perceptual encoding of the residual signal results in a substantial PESQ gain [2]. Furthermore, the final average PESQ scores for G728 and engineered coding method is 3.44 and 3.93 respectively. The resulting quality of speech coding is then comparable to the quality achieved with the ADPCM methods operating at similar bitrate. Unfortunately, the partial PESQ gain is negligible or even negative for segments of the residual signals containing transients. The reason for that is related to pre-echo distortions introduced by the employed perceptual module [12]. Those artifacts are expected to be reduced when the perceptual module dedicated to the speech signal would be employed instead of the Ogg Vorbis codec. It is worth mentioning that elimination of those artifacts is also possible if only the pure-voiced parts of the residual signal would be perceptually encoded. The results of conducted experiments confirm however, that additional encoding of the CELP

residual signal is an efficient technique for improving the quality of speech.

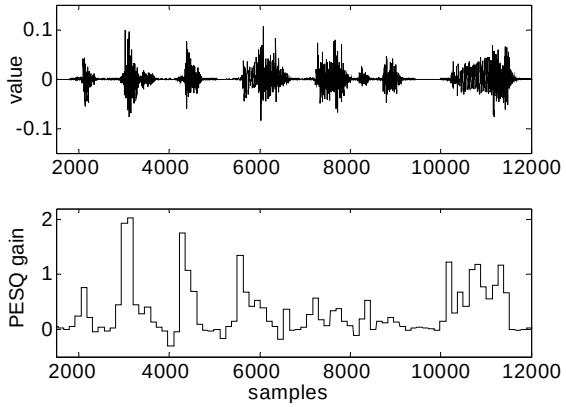

Fig. 7 Voiced part of the G728 residual signal along with corresponding partial PESQ gains function for entire speech signal

V. Speech Codec Employing Voiced/Unvoiced/Transient Segmentation

The drawback of the residual signal encoding is that it requires first local decoding of the CELP bit stream, and then the additional delay is introduced into the entire encoding process. Furthermore, conducted experiments have revealed that perceptual encoding of the speech signal transients is inefficient when the module operates in the low bit-rate mode. Therefore, the architecture of speech codec employing unvoiced/voiced/transients segmentation was investigated next. The diagram illustrating the proposed encoder architecture is shown in Fig. 8.

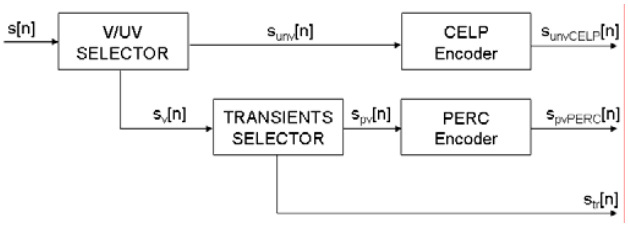

Fig. 8 Hybrid encoder architecture

In the first step, the input signal is split into unvoiced $s_{unv}[n]$, pure-voiced $s_{pv}[n]$ and transients $s_{tr}[n]$ parts. Further, each part is encoded employing an appropriate method. In practical experiments the G728 encoder (16 kbps) and the Ogg Vorbis perceptual module (14 kbps) were employed in order to encode the unvoiced and pure-voiced parts of speech signal respectively. As the purpose of the proposed experiment was to check if hybrid speech encoding results in quality improvement, the transients were not encoded with a use of any lossy method. It was found that about 40% of input signal frames were classified as unvoiced ones and about 7% as transient segments [2]. Thus, the overall estimated bitrate for unvoiced bit stream is about 6.4 kbps (0.4*16000) and 9 kbps (0.07*16*8000) in case of 16-bits PCM transients representation. Consequently, the examined encoder operates at average 29.6 kbps bitrate. Obviously, an appropriate method for transients coding would increase the efficiency of proposed codec architecture, and thus will be devised in a future work. It is expected that hybrid encoder would then operate at 24 kbps rate or even lower. In the decoder, the unvoiced and pure-voiced bit streams are decoded using G728 and Ogg decoders, respectively. Finally, these two parts are added together with the $s_{tr}[n]$ signal and in this way the entire speech signal is obtained.

The PESQ measurements were taken in order to evaluate the quality of speech samples encoded with a use of the hybrid method. In Fig. 9 comparison of partial PESQ scores obtained for a speech sample encoded with G726 ADPCM operating at 32 kbps and hybrid codec is presented.

Although it can be seen from Fig. 9 that partial PESQ scores are similar for that particular speech sample, the total average PESQ score obtained for male and female speech samples is slightly higher for hybrid codec. Fig. 10 presents the mean PESQ scores for various speech codecs architectures and also the results obtained for two codecs proposed in this paper. All of quality evaluation tests were performed with the use of the OPERA application [13].

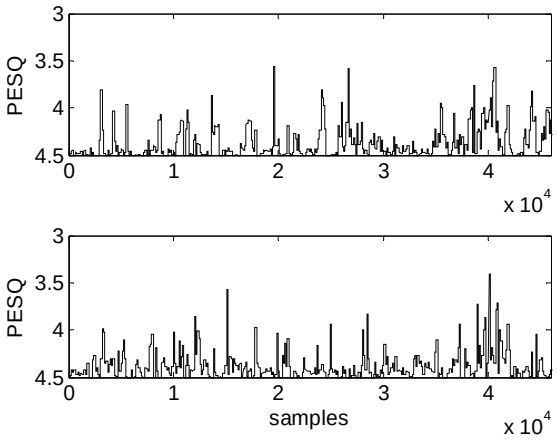

Fig. 9 Partial PESQ plots for G726 (upper) and hybrid codec (lower)

Comparing the PESQ scores for various codec architectures shown in Fig. 10 it is visible that in order to achieve similar quality of speech coding the hybrid codec requires lower average bitrate than the ADPCM codec. An additional efficiency improvement is expected after replacing the general audio perceptual module with the one dedicated to speech coding. Furthermore, in order to reduce the bitrate requirements it is also necessary to engineer an appropriate method for transients coding. It has to be mentioned that possible and simple solution to control the codec's properties is to scale the bitrate of CELP and perceptual module. Although the proposed method introduces a relatively high delay due to perceptual module involving, it allows also coding speech signal with a higher quality than other standardized methods.

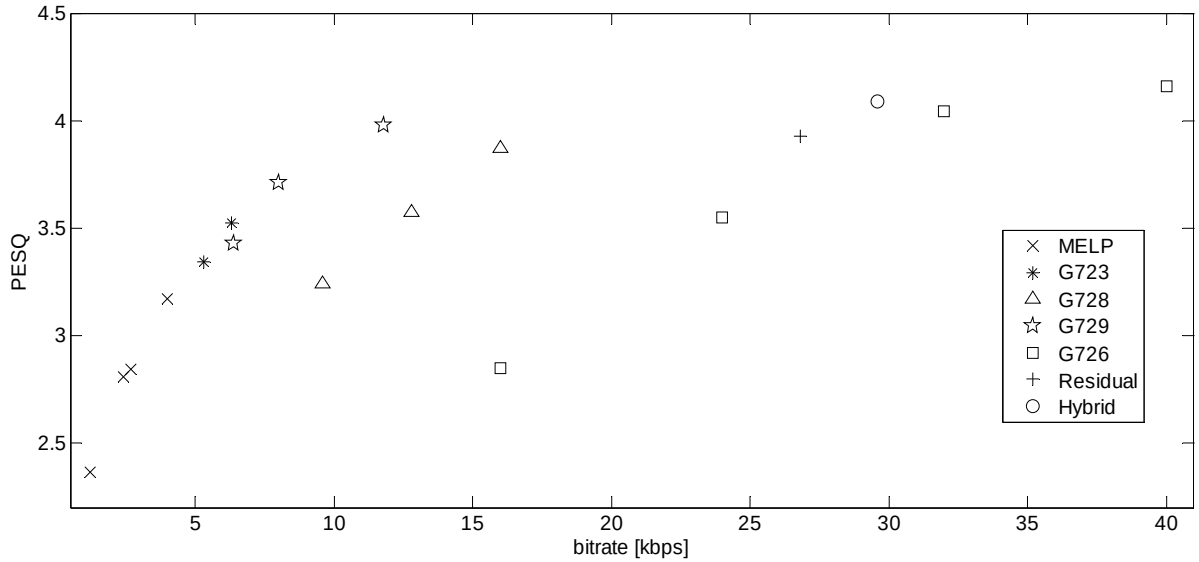

Fig. 10 Quality comparisons of various speech codecs architectures

VI. Conclusions

It can be seen that there is a wide variety of speech coders available, each one providing a different speech coding quality. In the past years the effort was made to develop the methods for low and middle bit-rate speech coding with satisfactory intelligibility. Hence, there is still a gap in terms of quality between middle bit-rate CELP codecs and high bit-rate ADPCM algorithms. In this paper two methods for incorporating the perceptual algorithm into parametric codec were proposed. The conducted experiments have revealed that perceptual encoding of the voiced part of CELP residual signal is the most straightforward method for improving the quality of speech coding. However, it has to be mentioned also, that in this approach only the quality of voiced parts of speech improved, and then the overall quality of speech signal remains limited. Hence, the second codec architecture employing voiced/unvoiced/transient segmentation of the entire speech signal was investigated. In turn it was found experimentally that perceptual encoding of the pure-voiced parts and transient states preserves high quality speech coding. In this approach, the unvoiced parts of the speech signal were encoded employing a parametric technique. An additional efficiency improvement is expected after replacing the general audio perceptual module with the one dedicated to speech coding. Furthermore, in order to reduce the bitrate requirements, it is also necessary to engineer an appropriate method for transients coding. Finally, one can conclude that the engineered hybrid codec algorithm allows for high quality speech coding, and further efficiency improvement is still possible to obtain after an additional optimization of the proposed algorithms.

Acknowledgment

Research funded by the Ministry of Education and Science within the Grant No. 3 T11D 004 28.

References

[1] Yang M., *Low bit rate speech coding*, IEEE Potentials, vol. 23, no. 4, pp. 32-36, 2004.
[2] Kulesza M., Szwoch G., Czyżewski A., *Improving signal quality in speech codec using hybrid perceptual-parametric algorithm*, Multimedia and Network Information Systems' 06, Wrocław, (submitted for publication).
[3] Ritz C. H., *Lossless wideband speech coding*, 10th International Conference on Speech Science and Technology, Sydney, Australia, December 2004.
[4] Dong H., Gibson J.D., *Structures for SNR scalable speech coding*, IEEE Transactions on speech and audio processing, (accepted and to appear) May 2006.
[5] Verma T.S., Levine S.N., Meng T.H., *Transient Modeling Synthesis: a flexible analysis/synthesis tool for transient signals*. International Computer Music Conference, Greece, 1997.
[6] Chu W.C., Speech Coding Algorithms. Foundation and Evolution of Standardized Coders, John Wiley & Sons, Hoboken 2003.
[7] Goldberg R., Riek L., *A Practical Handbook of Speech Coders*, CRC Press, Boca Raton 2000.
[8] Kliewer J., Mertins A., *Audio subband coding with improved representation of transient signal segments*, Proc IX European Signal Processing Conference (EUSICPO-98), Rhodes, Greece, September 1998, pp. 1245-1248.
[9] Babu V. S., Malot A. K., V. M. Vijayachandran V.M., Vinay M. K., *Transient Detection for Transform Domain Coders*, AES 116th Convention, Berlin, May 2004.
[10] ISO / IEC 14496-3:2001 Information technology - Generic coding of moving pictures and associated audio information: Part 3: Advanced Audio Coding (AAC).
[11] *OGG Vorbis Specification*: http://xiph.org/vorbis/
[12] Painter T., Spanias A., *Perceptual Coding of Digital Audio*, Proceedings of IEEE, vol. 88, pp. 451-513, April 2000.
[13] Opticom, *Opera your digital ear*, User manual, version 3.5, 2002.