A novel super-wideband embedded speech and audio codec based on ITU-T Recommendation G.729.1

Mao-shen JIA, Chang-chun BAO, Xin LIU, Rui LI
Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing 100124, China
E-mail: jiamaoshen@emails.bjut.edu.cn, baochch@bjut.edu.cn, liuxin0930@emails.bjut.edu.cn, lirui110@emails.bjut.edu.cn

***Abstract*—This paper proposes a multi-layer super-wideband embedded speech and audio coding algorithm extending bit rates from 36 to 64 kb/s on the basis of ITU-T Recommendation G.729.1 with a multi-stage coding structure. This codec consists of three embedded stages: G.729.1 wideband coding operating in the range from 8 to 32 kb/s, modified Modulated Lapped Transform (MLT) coding of the band (7-14 kHz) at 36, 40 & 48 kb/s and MDCT transform coding for wideband residual signal at 56 and 64 kb/s. In addition, some methods are proposed in transform coding according to perception significance. The objective and subjective listening tests show that this codec has good performance compared with reference codec.**

I. INTRODUCTION

Embedded coding is a method that the signal can be encoded in different bit-rates layers. Depending on the significance of the coding parameters, the bit-stream is generally divided into several layers. When the information in higher layers was lost in transmission, the lower layer can be independently decoded by the embedded decoder. In this bit-stream structure, the core layer contains the basic information providing a basic decoding quality. Each enhancement layers contains extra information and improves the performance of the decoder [1, 2]. With the bit rate increasing, the quality improvement of the reconstructed signal occurs obviously.

Embedded coding is absolutely necessary for the network communications and Web-based applications for mobile communication system. So ITU-T SG 16 began to study a new wideband speech coding standard called G.VBR in 2005, and it renamed as G.718 was standardized in June 2008. In July 2007, SG 16 agreed to develop a joint tool extending G.729.1 and G.VBR with super-wideband functionality, and several research institutions are participating in the work of standardization. The study in this paper is for competition of this super-wideband coding standard.

In the paper, a super-wideband embedded speech and audio coding algorithm is proposed. On the basis of the ITU-T Recommendation G.729.1, two embedded stages are added: MLT coding and wideband MDCT residuals advanced coding. It makes sure that the output bit-stream produced by the encoder is embedded, so this codec has more robustness in transmission. This codec can process signal with 32 kHz sampling rate and operate in the range from 8 up to 64kb/s.

The rest of this paper is organized as follows: The features of the codec are described in section 2. The encoder is described in section 3, while the decoder is presented in section 4. The results of objective/subjective listening tests and the conclusion are discussed in section 5 and 6 separately.

II. OVERVIEW OF MAIN CODEC

This codec is a multi-layer super-wideband embedded variable bit-rate speech and audio codec which can process signals sampled at 32 kHz with a bandwidth of 14 kHz in terms of scalable bit rates. The overall algorithmic delay of the codec is 50.9375ms. The contributions to this delay are 20 ms for the length of frame, 20ms for overlap-adding, 5 ms for LPC look-ahead, 2ms for re-sampling, and 3.9375 ms for analysis-synthesis QMF (Quadrature Mirror Filter-bank).

Five bit rates including 36, 40, 48, 56 & 64 kb/s are extended to G.729.1 in this codec. The information from 0 to 7 kHz is coded by the G.729.1 codec at bit rates from 8 to 32 kb/s, the frequency-domain information from 7 to 14 kHz is coded at 36, 40&48kb/s using modified MLT coding and the frequency-domain residual information from 0 to 7 kHz is coded at 56, 64kb/s.

III. DESCRIPTION OF THE ENCODER

The encoder is illustrated in Fig. 1. There are three operation modules in encoder: wideband G.729.1 coding module, MLT coding module, and wideband residual signal's MDCT coding module. There are 640 samples in each frame.

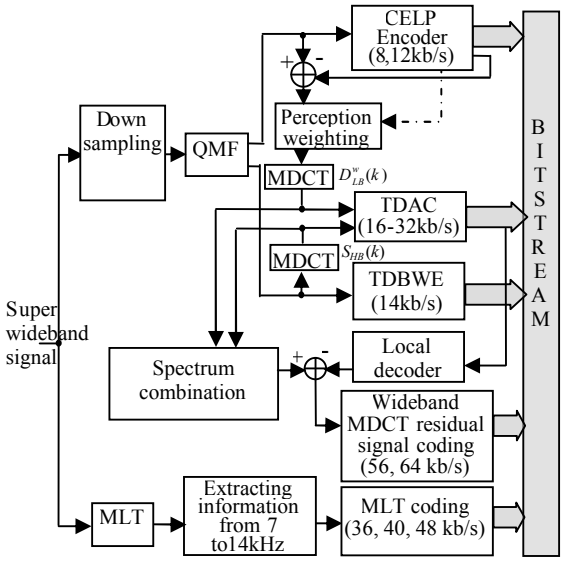

Fig. 1 Block diagram of the encoder

A. Down sampling

In order to make codec interoperable with the G.729.1, the supper-wideband input signal sampled at 32 kHz should be down-sampled before processing. The down-sampling is described as follows:

$$S_{wb}(n) = \sum h(i) \cdot S_{input}(2n - i) \quad (1)$$

where $h(n)$ is an 64-tapped low-pass FIR filter, $S_{input}(n)$ is the super-wideband input signal, and $S_{wb}(n)$ is the down-sampled signal sampled at 16kHz.

B. G.729.1 wideband encoding

The wideband signal $S_{wb}(n)$ is coded by G.729.1 codec at 12 different bit-rates from 8 to 32 kb/s.

By using a 64-coefficients analysis QMF, the signal $S_{wb}(n)$ is decomposed into two sub-bands. The lower band which includes the information from 0 to 4 kHz is encoded by a cascade CELP coder at 8, 12 kb/s. The higher band which includes the information from 4 to 8 kHz is spectrally folded and encoded by parametric time-domain bandwidth extension (TDBWE) at 14 kb/s. The lower band CELP residual signal and the higher band original signal are jointly encoded at bit rates from 16 to 32 kb/s by the time-domain aliasing cancellation (TDAC) encoder. [3, 4]

C. MLT coding (36, 40, 48kb/s)

In 36, 40, 48kb/s bit-rate layers, this paper proposes a modified MLT transform coding which is an improved method of ITU-T Recommendation G.722.1.

Firstly, the supper-wideband input signal $S_{input}(n)$ is transformed to frequency domain by Modulated Lapped Transform; the definition of the MLT is given by:

$$m_{lt}(m) = \sum_{n=0}^{1279} \sqrt{\frac{2}{640}} \sin\left(\frac{\pi}{1280}(n+0.5)\right) \cos\left(\frac{\pi}{640}(n-319.5)(m+0.5)\right) S_{input}(n) \quad (2)$$

where $0 \le m < 640$, $S_{input}(n)$ includes 1280 samples which are from previous frame and current frame, $m_{lt}(m)$ are the MLT coefficients. It means that the most recent 1280 samples are transformed into 640 MLT coefficients for each frame. 280 MLT coefficients representing frequencies from 7 to 14 kHz are extracted for processing. The extracted coefficients are coded in 36,40,48kb/s layers, so the maximum bit rate for these extracted coefficients is 16kb/s, i.e. 320 bits per frame are allocated for MLT coefficients.

Secondly, these coefficients are divided into 14 regions; each region represents a bandwidth of 500 Hz and includes twenty coefficients. The region power for each of the 14 regions is defined as the root-mean-square value of the MLT coefficients in that region, and it is computed as:

$$r_{ms}(k) = \sqrt{\frac{1}{20}\sum_{n=0}^{19} m_{lt}^2(20k+n)} \quad (3)$$

where k is the number of region. The region power is quantized in the logarithmic domain, and the indexes are encoded by Huffman codes.

Thirdly, depending on each region's index and the number of available bits used for quantization of MLT coefficients, 16 categorizations can be produced by the categorization procedure. In the categorization, every region will be assigned a category; in different category, the quantization and coding parameters and bit allocation for MLT coefficients is different. There are eight categories: 0, 1,…, 7. Each categorization consists of a set of 14 "category" assignments, it means that each region have one category. In different categorization, the 14 category assignments are different. The parameters used for encoding MLT coefficients are determined by the categorization. 16 categorizations are ordered according to their expected number of bits. Categorization 0 has the largest expected number of bits and categorization 15 has the smallest bits number [5, 6]. Only the categorization whose expected number of bits is the nearest to the available number of bits is selected to quantize the MLT coefficients, and the number of the categorization is described by the 4-bit categorization control bits.

Fourthly, in the region which assigns categories 0 through 6, MLT coefficients are separated into sign and magnitude information. The magnitude of the MLT coefficients are normalized by the quantized region power and then quantized with scalar quantizer. The quantization indices are combined into vector indices and encoded by Huffman code. The sign bits are directly represented by 0 or 1. The coefficients in the region which assigns category 7 are always set to "0".

From psychoacoustic principles, we know that the regions with bigger power are more important than others. To improve the efficiency of MLT coding, we propose here to include frequency-domain perception significance order of MLT spectrum. Such ordering method can modify the region sequence in transmission originally in the natural frequency ascending order to that in the region power descending order. Therefore, the regions are sorted according to the quantized region power $r_{ms\_Q}(k)$ in descending order. The ordered index sequence is represented as $R_{ind}(k)$, k=1,2,.., 14, where region $R_{ind}(1)$ has the biggest power quantization value and region $R_{ind}(14)$ has the smallest power quantization value. This imply that region $R_{ind}(1)$ is the most important and region $R_{ind}(14)$ is the least important in all regions.

According to the sequence $R_{ind}(k)$, k=1,2,…, 14, the MLT code bits are written into bit-stream. The bits from important regions are firstly written, and the bits from sub-important regions are written later. It makes sure that the bits in low bit-rates layer are more important than those in high layers.

In reference [2], we have shown the improved performance of embedded coding with significance order approach. Compared with G.722.1 codec, the MOS is averagely increased by 0.1 at the truncated bit-rates, 8, 16 and 24 kb/s.

The 36, 40, 48kb/s layer are coded together, and the bit stream is transmitted in three parts: region power code bits (variable number of bits), categorization control bits (4 bits), and the MLT code bits (variable number of bits). The structure of bit stream is shown in Fig. 2.

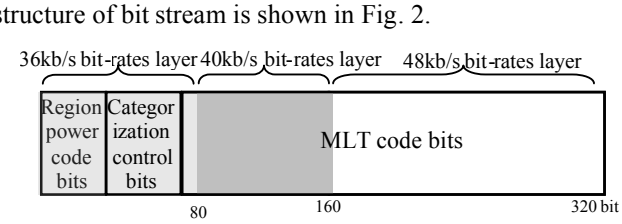

Fig. 2 Structure of bit-stream in 36, 40, 48kb/s layers

From Fig. 2, we can find that region power code bits, categorization control bits and some MLT code bits of important regions are in 36 kb/s layer, MLT code bits of other important regions are in 40kb/s layer, and the MLT code bits of sub-important regions are in 48kb/s layer.

D. Advanced coding of wideband MDCT residual signal (56, 64 kb/s)

In this coding module, the wideband MDCT residuals of G.729.1 coding module is fine quantized at 56 and 64 kb/s, with the advanced method which is used for coding MLT coefficients in section C. After this fine quantization in the bandwidth from 0 to 7 kHz, the quantization quality of wideband MDCT coefficients is improved in evidence and the subject reconstructed quality become more clear in the informal listening test.

The target signal of this module is a difference between two MDCT coefficients. One is from G.729.1 wideband coding module and another one is from original input signal. Therefore, it needs local synthesis of TDAC. Because TDAC's target signal is formed by concatenating the MDCT domain information of residual signal from 0 to 4 kHz ($D_{LB}^{w}(k)$,k=0,1,…,159) and the information of original signal from 4 to 7 kHz ($S_{HB}(k)$, k=0,1,…,119)[4], MDCT residuals are obtained from two different bands as follows:

$$\begin{cases} M_{LB}(k) = D_{LB}^{w}(k) - \hat{D}_{LB}^{w}(k) \\ M_{HB}(k) = S_{HB}(k) - \hat{S}_{HB}(k) \end{cases} \tag{4}$$

where $\hat{D}_{LB}^{w}(k)$ is the quantized MDCT coefficients of the local synthesis residuals from 0 to 4 kHz, and $\hat{S}_{HB}(k)$ is the MDCT coefficients of original signal from 4 to 7 kHz.

Next, $M_{LB}(k)$ and $M_{HB}(k)$ are jointly encoded by the advanced method mentioned in section C. The region power code bits, categorization control bits, and important regions code bits of MDCT residual coefficients are encoded in 56 kb/s layer, and the sub-important regions code bits of MDCT residual coefficients in 64 kb/s layer.

IV. DESCRIPTION OF THE DECODER

The decoder is illustrated in Fig. 3.

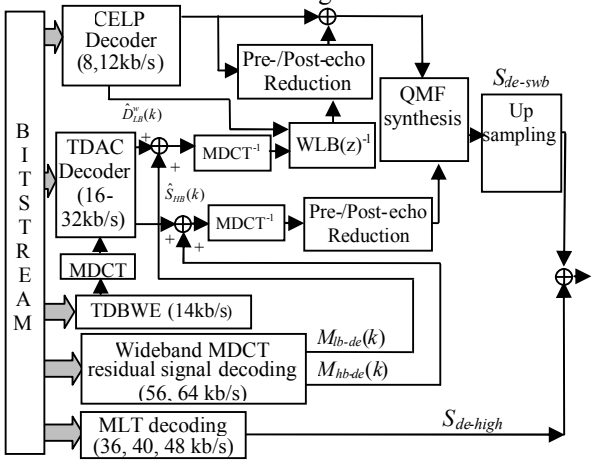

Fig. 3 Block diagram of the decoder

Same as the encoder, there are three operation modules: G.729.1 wideband decoding module, MLT decoding module, and wideband MDCT residual signal coding module. The decoder can receive any of the supported bit rates from 8 up to 64 kb/s. According to the received bit rates $R_{bit\text{-}rates}$, the decoder executes relevant operation.

If $R_{bit\text{-}rates}$≤32kb/s, the received bit-stream is processed by G.729.1 codec, and the decoder reconstructs the wide-band signal $S_{de\text{-}wb}$ at corresponding bit rates. After up-sampling, the super-wideband signal $S_{de\text{-}swb}$ sampled at 32 kHz with an effective bandwidth of 7 kHz is obtained.

If $R_{bit\text{-}rates}$=36kb/s, the layers under 36kb/s is decoded by the G.729.1. The bit-stream of 36kb/s layer consists of three parts: region power code bits, categorization control bits, and MLT code bits of partial important regions. The quantized region power sequence $\hat{r}_{ms\_Q}(k), k=1,2,...,14$ is decoded by Huffman codes. After ordering the sequence $\hat{r}_{ms\_Q}(k)$, the significance index sequence of regions $R_{ind}(k)$, k=1,2,.., 14 is obtained, and it is the same as the sequence in encoder. From this, we can know the regions whose MLT code bits is in 36kb/s layer. Depending on region power quantization and the number of available bits used for quantization of MLT coefficients, the 16 possible categorizations are reconstructed at the decoder with the same categorization procedure. The categorization which is used at the encoder can be indicated by the 4 categorization control bits.

In the important regions whose MLT code bits are in 36kb/s layer, the normalized and quantized MLT coefficients are recovered from the MLT vector indices according to the categorization. So MLT coefficient's amplitude is reconstructed as the product of region power and quantized normalized MLT coefficients. If some coefficient is non-zero value, it should have its sign set according to the sign bit [6].

In the sub-important region whose MLT code bits are not in 36 kb/s layer, the MLT coefficients are reconstructed by noise-filling as follows:

$$m_{lt}(i) = s_{ign\_random} \times r_{ms\_q}(r) \times \beta \tag{5}$$

where s_{ign_random} is the random sign, $r_{ms\_q}(r)$ is the quantized region power, β is the attenuation factor and $m_{lt}(i)$ is the reconstructed MLT coefficients. This operation is also done in the region that is assigned to category 7. So the MLT coefficients from 7 kHz to 14 kHz can be recovered.

If $R_{bit\text{-}rates}$=40kb/s or $R_{bit\text{-}rates}$=48kb/s, the bit-stream includes more MLT code bits, so more coefficients can be accurately decoded by the received bit-stream. In the region whose MLT code bits are not received in decoder, the MLT coefficients are also reconstructed by noise-filling.

When $R_{bit\text{-}rates}$=36kb/s, 40kb/s, or 48kb/s, the decoded MLT coefficients represent frequencies from 7 to 14 kHz. Whereas the 360 MLT coefficients representing frequencies 0~7 kHz and above 14 kHz are set to "0". Thus the reconstructed 640 MLT coefficients are converted into time domain samples $S_{de\text{-}high}$ by inverse MLT (IMLT). By computing the sum of $S_{de\text{-}high}$ and $S_{de\text{-}swb}$, the decoded signal is reconstructed.

If $R_{bit\text{-}rates}$=56kb/s or $R_{bit\text{-}rates}$=64kb/s, the layers under 56kb/s is decoded by the lower bit-rates decoder, and the 56, 64 kb/s bit-rates layer is decoded by the algorithm which is

used in 36, 40, 48kb/s layers. If $R_{bit\text{-}rates}$=64kb/s, all regions' coefficients can be decoded from the received bit-stream. If $R_{bit\text{-}rates}$=56kb/s, the MDCT residual coefficients whose code bits are not received in decoder are also reconstructed by noise-filling. So the MDCT residual coefficients $M_{dct\text{-}de}(k)$ representing frequencies from 0 to 7 kHz can be decoded. $M_{dct\text{-}de}(k)$ is divided into 2 parts: $M_{lb\text{-}de}(k)$ and $M_{hb\text{-}de}(k)$, representing frequencies 0~4 kHz and 4~7 kHz separately.

The modified MDCT coefficients $S_{HB\_DE}(k)$ and $D^{w}_{LB\_DE}(k)$ is produced by the operation as follows:

$$\begin{cases} D^{w}_{LB\_DE}(k) = M_{lb-de}(k) + \hat{D}^{w}_{LB}(k) \\ S_{HB\_DE}(k) = M_{hb-de}(k) + \hat{S}_{HB}(k) \end{cases} \quad (6)$$

where $\hat{S}_{HB}(k)$, $\hat{D}^{w}_{LB}(k)$ are the TDAC decoded coefficients. Then $\hat{S}_{HB}(k)$ and $\hat{D}^{w}_{LB}(k)$ are replaced by $S_{HB\_DE}(k)$ and $D^{w}_{LB\_DE}(k)$ to perform inverse MDCT in G.729.1 decoding.

Depending on bit stream of G.729.1, the decoded signal $S_{de\text{-}swb}$ obtained from different bit rate layers is different. When the third module is activated ($R_{bit\text{-}rates}$=56kb/s or $R_{bit\text{-}rates}$=64kb/s), $S_{de\text{-}swb}$ should be modified with decoded MDCT residual coefficients, then the decoded signal is the sum of the second module decoded signal $S_{de\text{-}high}$ and modified $S_{de\text{-}swb}$.

V. PERFORMANCE EVALUATION

Based on G.729.1 there are 5 bit rates added in proposed codec: 36, 40, 48, 56 & 64 kb/s. Terms of Reference (ToR) which is request by ITU-T for the candidate super-wideband codec is that: the quality of candidate codec at 36, 40kb/s is not worse than G.722.1C at 24kb/s, the quality of candidate codec at 48, 56 kb/s is not worse than G.722.1C at 32kb/s, and the quality of candidate codec at 64 kb/s is not worse than G.722.1C at 48 kb/s. ITU-T's TOR is taken as a reference for tests in this paper. The test data from the MPEG database include 8 audio segments and 4 speech sentences (in English). The sampling rate is 32 kHz and the nominal input level is -26 dB with respect to the OVL point.

The objective performance test of the proposed codec and reference codec G.722.1C is compared by Objective Difference Grade (ODG) scores by using Perceptual Evaluation of Audio Quality of ITU-R BS.1387 [7, 8]. The PEAQ test compares the perceptual difference between the processed signal and the original one. The grading scale of ODG ranges from -4 (very annoying) to 0 (imperceptible difference). The average ODG results are given by table I.

TABLE I
ODG RESULTS OF THE TWO CODERS

Proposed codec (Bit rates)	ODG	G.722.1C (Bit rates)	ODG
36kb/s	-3.5263	24kb/s	-3.6995
40kb/s	-3.44673	24kb/s	-3.6995
48kb/s	-3.3263	32kb/s	-3.592
56kb/s	-3.22	32kb/s	-3.592
64kb/s	-3.0749	48kb/s	-3.3063

From table I, we can see that the objective performance of proposed codec is better than the reference codec at 36, 40, 48, 56 and 64 kb/s.

In order to further evaluate our codec's performance, an informal subjective A/B listening test is finished by twenty listeners (10 males and 10 females) through high quality headphones. The results are listed in table II. The listening test results show that proposed codec provides better subjective quality at 40, 48, 56 kb/s and comparable quality to reference codec at 36, 64 kb/s.

TABLE II
THE SUBJECTIVE A/B LISTENING TEST RESULTS

Bit rates	Prefer proposed codec	Prefer G.722.1C	No preference
36kb/s	35.2%	39.2%	26.6%
40kb/s	42.3%	20.5%	38.2%
48kb/s	34.4%	18.3%	48.3%
56kb/s	30.5%	16.2%	54.3%
64kb/s	28.6%	29.2%	43.2%

VI. CONCLUSION

Based on ITU-T Recommendation G.729.1 and modify MLT transform coding algorithm, this paper describes a multi-layers super-wideband embedded speech and audio codec. It is an extension of G.729.1. The results of objective and subjective listening tests show that this codec has good performance as the ITU-T reference codec. Compared to other un-embedded codec, this codec has more robustness and very high audio quality in the transmission and has some potential applications in the area of mobile audio, IP phone, audio/video conferencing and 3G & 4G mobile communications.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China under Grant 60872027, Beijing Natural Science Foundation under Grant 4082006 and Beijing University of Technology Ph.D. Innovation Projects bcx-2009-014.

REFERENCES

[1] Balázs Kövesi et al., "A scalable speech and audio coding scheme with continuous bitrate flexibility," in Proc. ICASSP, May 2004, vol.1, pp. 273-276.

[2] Mao-shen JIA, Chang-chun BAO et al. "A Novel Embedded Speech and Audio Codec Based on ITU-T Recommendation G.722.1," ICSP 2008, October 26-29, 2008，pp. 522-525

[3] ITU-T Rec. G.729.1, "An 8-32 kbit/s scalable wideband coder bit-stream interoperable with G.729," May 2006.

[4] Ragot. S et al., "ITU-T G.729.1: An 8-32 kbit/s Scalable Coder Interoperable with G.729 for Wideband Telephony and Voice over IP," in Proc. ICASSP, April 2007, vol. 4, pp.529-532.

[5] ITU-T Rec. G.722.1, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," September 1999.

[6] Minjie Xie et al., "ITU-T G.722.1 Annex C: A New Low-Complexity 14 kHZ Audio Coding Standard," in Proc. ICASSP, May 2006, vol. 5, pp.173-176.

[7] ITU-R BS.1387, "Method for objective measurements of perceived audio quality," 2001

[8] F. Baumgarte, A. Lerch, "Implementation of Recommendation ITU-R BS.1387, Delayed Contribution," Document 6Q/18-E, Feb, 2001.