International Journal of Speech Technology (2022) 25:219–230
https://doi.org/10.1007/s10772-021-09949-2

Information hiding in proposed 10.6 kbps CS-ACELP based speech codec using Quantization Index Modulation

Nikunj Tahilramani[1] · Ninad Bhatt[2]

Received: 1 February 2021 / Accepted: 14 November 2021 / Published online: 3 December 2021

Abstract
This paper aims to propose information hiding with the variant like steganography and watermarking in proposed 10.6 kbps Conjugate Structure-Algebraic Code Excited Linear Prediction (CS-ACELP) speech codec. Proposed work makes use of Dither Modulation-Quantization Index Modulation technique to incorporate steganography or watermarking for information hiding in the excitation codebook code vector of proposed 10.6 kbps CS-ACELP based speech codec. Codebook partition and label assignment approach is explored in proposed coder in order to create room of 10 bits per frame for steganographic data transmission. Joint source coding and data hiding approach is adopted for steganographic or watermarked data transmission. Performance of the proposed approach is evaluated with different objective and subjective parameters in terms of tables and graphs. Information hiding capacity is demonstrated with different parameters like the watermark to signal ratio, hiding capacity and embedding capacity in terms of percentage. Moreover, the results of subjective and objective parameters of the proposed algorithm are analysed by computing population mean of 99% confidence interval to prove the consistency of it.

Keywords Steganography · Watermarking · CS-ACELP · Excitation codebook code vector · Embedding capacity · Hiding capacity · Population mean · Confidence interval

1 Introduction

Code Excited Linear Prediction (CELP) is a class of hybrid coder which combines the features of both source coders and waveform coders. Hybrid coders always employ analysis by synthesis process in order to derive code parameters. In analysis by synthesis model, no prior knowledge of voiced/unvoiced or pitch period is needed. In CELP, residuals after short term and long-term prediction becomes noise like and it's assumed that the residual can be modelled by zero mean Gaussian process with slowly varying power spectrum. As a result, the excitation frame may be vector quantized using a large stochastic codebook.

Speech coding standards are established by various standards organizations such as ITU-T (International Telecommunication Union), Telecommunication standardization sector (CCITT), European Telecommunication Standards Institute (ETSI) etc. Nowadays it is very much important to develop low bit rate speech coding systems for voice storage such as voice mail and voice driven directory enquiries. The ultimate goal is to design a speech codec with good perceptual quality of digital mobile communication and other VoIP based application which indeed dominating heavily in the era of smart phone based communication world. VoIP based communication requires end to end secure speech communication to maintain confidentiality in the communication world of internet. The information hiding techniques like Steganography, watermarking and cryptography are the classified techniques for maintaining secrecy of the voice traffic as well as data traffic information between end devices. In the proposed work, information hiding takes place using Dither Modulation-Quantization Index Modulation, which is among a class of potential candidates of novel and emerging

✉ Nikunj Tahilramani
nikunjvtec@gmail.com

Ninad Bhatt
bhattninad@gmail.com

[1] Information & Communication Technology Department, Adani Institute of Infrastructure Engineering, Ahmedabad, Gujarat, India

[2] Electronics & Communication Engineering Department, C. K. Pithawalla College of Engineering & Technology, Surat, Gujarat, India

techniques of information hiding in the near future of fifth generation (5G) communication network.

This paper is systematized as follows. In Segment 2, Excitation sequence generation in the CS-ACELP based speech codec is depicted. In Segment 3, Transmission of codevector indices and its sign magnitude of stochastic code structure of 8 kbps CS-ACELP (Salami et al. 1998) speech codec are introduced. Segment 4 touches upon replacement of excitation codebook structure of 8 kbps CS-ACELP speech codec with the standard extended 11.8 kbps CS-ACELP speech codec. Segment 5 deals with the role of search engine for determination of optimized codevector. In Sect. 6, novel approach of codebook partition and label assignment approach for transmission of indices of excitation codevector is discussed. In Segment 7 and Segment 8, information hiding in proposed CS-ACELP speech codec with Quantization Index Modulation is elucidated. In Sect. 9, different objective (Hu and Loizou 2008) and subjective quality (ITU-T Recommendation 2003) assessment parameters of speech signal is listed. In Segment 10, performance assessment of proposed 10.6 kbps CS-ACELP speech codec and its comparative analysis approach with information hiding is demonstrated with the outcomes of different objective and subjective parameters in terms of set of tables and graphs. At last, the concluding remarks and summary are specified in Segment 11.

In the subsequent sections, 8 Kbps speech codec is defined as legacy speech codec, extended G.729E working at 11.8 Kbps is designated as standard speech codec, modified 11.8 Kbps which is proposed 11.6 Kbps is designated as modified extended speech codec and 10.6 Kbps speech codec is notified as proposed speech codec.

2 Excitation sequence generation in conjugate structure-algebraic code excited linear prediction based speech codec

The generation of excitation sequence is shown in the Fig. 1. The segmented speech frames are first pre-processed and then short-term analysis is performed in terms of Linear Predictive coding over present 80 speech samples which results into 10 LP coefficients per frame (ITU-T Recommendation, 2007). The resultant LP coefficients and segmented speech samples are given input to the weighting filter to avoid the effect of noise generated at the synthesis stage due to noise located in lower energy regions. According to the analysis by synthesis principle, for each sub frame time slots, the output of the weighted error is deducted from the output of the filter impulse response which is a grouping of error weighting filter and short-term synthesis filter (Tahilramani

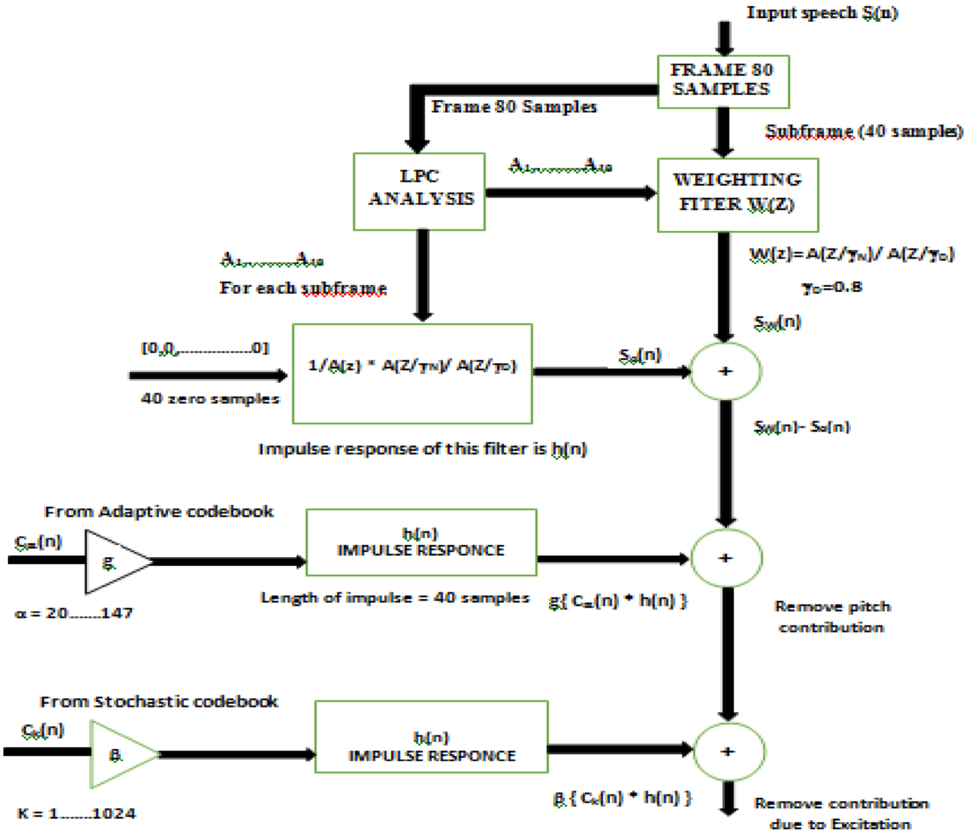

Fig. 1 Excitation sequence generation of CS-ACELP based speech codec (Tahilramani & Bhatt, 2017)

Table 1 Bit allocation of standard 8 kbps CS-ACELP based speech codec (ITU-T Recommendation, 2007)

Speech parameters	Sub frame 1	Sub frame 2	Total bits/frame
Pitch-delay parity	1		1
Fixed-codebook index	13	13	26
Adaptive-codebook delay	8	5	13
LSP			18
Codebook gains (stage 1)	3	3	6
Fixed-codebook sign	4	4	8
Codebook gains (stage 2)	4	4	8
Total			80

Table 2 Stochastic excitation code-structure of CS-ACELP based 8 kbps speech coder (ITU-T Recommendation, 2007)

Pulse	Sign	Positions
i_0	$s_0: \pm 1$	m_0: 0, 5, 10, 15, 20, 25, 30, 35
i_1	$s_1: \pm 1$	m_1: 1, 6, 11, 16, 21, 26, 31, 36
i_2	$s_2: \pm 1$	m_2: 2, 7, 12, 17, 22, 27, 32, 37
i_3	$s_3: \pm 1$	m_3: 3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

& Bhatt, 2017). For long term analysis, contribution of pitch is subtracted from the output by subtracting the combination of impulse response and adaptive codebook from it. The output without pitch contribution is again subtracted from the combination of the impulse response of the filter and the stochastic or excitation codebook to produce short term residuals which is given as feedback to excitation input to optimize the excitation.

3 Transmission of non-zero pulse position and its sign magnitude in excitation codebook structure of CS-ACELP based speech codec

In the proposed work, one of the speech codec based on CS-ACELP which is a variant of CELP is taken as a base coder which is standardised by ITU-T and operating at a bit rate of 8 kbps (G.729) (ITU-T Recommendation, 2007). In base codec, each sample of digitized speech is represented as 16 bits with sampling frequency of 16,000 samples/second. The time slot 1 frame of 10 ms with 80 bits per frame transmission leads to the bit rate of 8 kbps. The Bit allocation of the standard CS-ACELP working at 8 kbps is highlighted in Table 1.

In base coder, the excitation frame is modelled by a Gaussian vector chosen from a large Gaussian codebook which is nothing but a residual signal after short term and long-term prediction. The excitation codebook structure of base coder (ITU-T Recommendation, 2007) is demonstrated in Table 2. As index of the selected pulse is transmitted, each pulse position is coded with 3 bits as there are 8 pulse positions in each track. Two pulse positions are selected after recursive search procedure and the position index and sign of the respective pulse is transmitted per sub-frame of 40 samples. As per the analysis by synthesis principle, 10 pulse positions are transmitted per 5 ms slot of sub-frame which requires to be coded with 30 bits.

As per the conventional way, 10 sign bits are required to be transmitted per sub-frame for 10 pulse positions. To reduce the number of bits transmitted per sub-frame for the sign bit parameter, sign bit of only one pulse is transmitted with certain assumptions so that the sign of the other pulse position can be judged based on the sign of the one pulse transmitted per track. If sign of both pulse positions belonging to same track are same then index of the smallest among two pulse position is transmitted first with 1 sign bit of the respective smallest pulse is transmitted. In other case index of the biggest among two pulse positions is transmitted first with its respective sign bit. With above two cases, the sign of the remaining pulse which is not transmitted to the decoder can be judged accordingly.

4 Replacement of fixed codebook structure of legacy speech codec with forward mode excitation codebook structure of standard speech codec

Excitation codebook structure of ITU-T standardised speech codec is having 4 tracks (ITU-T Recommendation, 2007) with first three tracks are having 8 pulse position while 4th track is having pulse position. Final codevector is having contribution from four non-zero pulses from four different tracks which are selected by focused search approach (Bernard, 2005). Due to 16 pulse positions in final track of excitation codebook structure (ITU-T Recommendation, 2007), the number of searches required to determine final excitation codevector increases as search engine needs to consider each combination twice for finding out the best combination; which determines the final excitation codevector. The recursive procedure for finding out the excitation codevector commence from different track each time by assigning four different pulses starting from four different tracks at each stage of focused search approach.

To reduce the number of searches in proposed speech codec, the excitation codebook structure of a base codec (ITU-T Recommendation, 2007) is replaced with the forward mode excitation codebook structure of another variant of standard speech codec, which is extended G.729

(G.729E) ITU-T standardised speech codec working at 11. 8 kbps (ITU-T Recommendation, 2007) consist of five tracks from those each track consists of two non-zero pulse positions per track in excitation codebook structure. To measure the contribution of each pulse to determine optimized final excitation codevector, least significant pulse replacement approach (Lee & Kim 2010) is applied on different combinations of pulse position, which actually reduces the number of searches required to form a final codevector, which is depicted in Sect. 3. The excitation codebook structure and bit allocation of ITU-T standardised speech codec is shown in Tables 3 and 4.

From the Table 1, it can be observed that the fixed codebook index and sign parameter in legacy speech codec requires 34 bits per segment slot (10 ms) for transmission of excitation codevector pulse positions indices and their corresponding sign bit (ITU-T Recommendation, 2007) (80 samples). Since standard speech codec uses two dissimilar bits stochastic codebook structure at encoder and decoder, it needs 70 bits per segment for transmission of fixed codebook indices with its respective signs of pulse positions (ITU-T Recommendation, 2007) (10 ms, 80 samples) in forward mode (Table 4). The bit allocation of Modified extended (11.6 kbps) speech codec is shown in Table 5.

5 Optimized exhaustive search engine for excitation codevector in CS-ACELP based speech codec

Modified extended speech codec uses forward mode stochastic codebook structure of extended G.729E for determining best optimized excitation codevector using least significant pulse replacement approach. The excitation codebook structure consists of five tracks with each track consist of eight different pulse positions. The final excitation codevector consist of ten pulse positions having two pulse positions contribution from each track consist of sign pulse magnitude ±1.

Table 3 Excitation codebook structure of standard (11.8 Kbps) speech codec (ITU-T Recommendation, 2007)

Track	Pulses	Signs	Positions
1	$m0$, $m1$	$s_0, s_1: \pm 1$	0, 5, 10, 15, 20, 25, 30, 35
2	$m2$, $m3$	$s_2, s_3: \pm 1$	1, 6, 11, 16, 21, 26, 31, 36
3	$m4$, $m5$	$s_4, s_5: \pm 1$	2, 7, 12, 17, 22, 27, 32, 37
4	$m6$, $m7$	$s_6, s_7: \pm 1$	3, 8, 13, 18, 23, 28, 33, 38
5	$m8$, $m9$	$s_8, s_9: \pm 1$	4, 9, 14, 19, 24, 29, 34, 39

Table 4 Bit allocation of standard (11.8 Kbps) speech codec (ITU-T Recommendation, 2007)

LP mode indication bit	G.729E 11.8 Kbit/s	
	1 + 1(parity)	
	Backward	Forward
EXC codes (1st/2nd sub-frame)	44/44	35/35
Gains (LTP + EXC) (1st/2nd sub-frame)	7/7	7/7
LTP delay (1st/2nd sub-frame)	8 + 1(parity)/5	8 + 1(parity)/5
LP filter	0	18
Total	118	118

It requires 35 bits for transmission of index of excitation codevector pulse position and sign of respective pulse position (±1) per sub frame (40 samples).

5.1 Search procedure for determination of excitation codevector

Exhaustive search procedure starts with determination of initial codevector by selecting the first two largest pulse positions on the basis of magnitude d(n) which is called as correlation vector (Bernard, 2005; ITU-T Recommendation, 2007) (given in Eq. (1)) from every track. The second stage codevector is investigated by substituting pulse position of two best selected pulse position of primary codevector with the supplementary pulse position of that particular track one by one. The above iterative procedure is repeated for each track. The combination of the pulse position which maximize the value of (Eq. (1)) $\mathrm{Q_k}$ from each track, is selected as next stage codevector. The above computation requires calculation of 12 $\mathrm{Q_k}$ values per track and 60 $\mathrm{Q_k}$ values for 5 tracks.

$$\mathrm{d(n)} = \sum_{\mathrm{i=n}}^{\mathrm{M-1}} \mathrm{x_2(i)h}(i - \mathrm{n}), \quad \mathrm{i} = 0, \ldots ., \mathrm{M} \tag{1}$$

Table 5 Bit allocation of modified extended (11.6 kbps) speech codec

Parameter	Subframe 1	Subframe 2	Total bits/frame
Fixed-codebook index	30	30	60
Pitch-delay parity	1		1
Fixed-codebook sign	5	5	10
Codebook gains (stage 1)	3	3	6
Adaptive-codebook delay	8	5	13
Codebook gains (stage 2)	4	4	8
LSP			18
Total			116

$$\max_k Q_k = \max_k \frac{C_k^2}{E_k} = \max_k \frac{\left(d^t c_k\right)^2}{c_k^t \emptyset_{c_k}}$$
$$= \frac{\left(\sum_{j=0}^{M-1} s_j d\left(m_j\right)\right)^2}{\sum_{j=0}^{M-1} \emptyset\left(m_j, m_j\right) + 2\sum_{i=0}^{M-2}\sum_{j=i+1}^{M-1} s_i s_j \emptyset\left(m_i - m_j\right)} \quad (2)$$

where M denotes the total number of tracks in a sub frame analysis.

C_k is described as a Kth codebook vector and t represents a transposed matrix and PHI matrix is described as (ITU-T Recommendation, 2007):

$$\emptyset(i,j) = \sum_{n=j}^{M-1} h(n-i)h(n-j), j = i, \dots M \quad (3)$$

where $x_2(n)$ is target signal for the fixed codebook and h(n) indicates the impulse response of a synthesis filter.

Numerator and denominator of Eq. (2) is represented as (ITU-T Recommendation, 2007):

$$C = \sum_{i=0}^{N_p-1} \mathrm{sign}\{d(i)\}d\left(m_i\right) \quad (4)$$

$$E = \sum_{i=0}^{N_p-1} \emptyset\left(m_i, m_j\right) + 2\sum_{i=0}^{N_p-2}\sum_{j=i+1}^{N_p-1} \mathrm{sign}\{d(i)\}\mathrm{sign}(d(j)\}\emptyset\left(m_i, m_j\right) \quad (5)$$

Number of pulses in sub-frame is described as Np and m denotes a position of ith pulse.

6 Novel approach of codebook partition and label assignment for transmission of position indices and sign magnitude of non-zero pulses with reduced number of bits

The proposed approach of codebook partition and label assignment, initially the excitation codebook structure is divided into two equal parts with each partition is having four pulse position as each track in modified extended speech codec is having eight pulse positions. As each partition contains four pulse positions, pulse positions of each partitions can be assigned a label with two-bit combination instead of three-bit combination. Pulse positions from partition one is labelled in ascending order as 00, 01, 10 and 11 while in partition two it is labelled in descending order as 11, 10, 01 and 00.

The sign bit for both the pulse positions belonging to same track are transmitted with different configuration. If the selected two pulses belong to the same track then the sign magnitude of corresponding two pulses are transmitted with two in one cell and same remains true for two possible cases where the selected pulses either belong to partition one or partition two. In third possible case where both the pulse positions belong to different track then the magnitude of corresponding two pulse positions are transmitted one in one cell. Due to this novel way of transmission of sign magnitude pulses, the decoder of proposed CS-ACELP based speech codec checks the cell format of the bits of the sign magnitude pulses to determines the position of the indices of the selected pulse positions in excitation codebook structure in the two possible cases where either two pulse positions belong to same track or different track. If the selected pulse positions belong to same track then partition number is determined using their binary label assignments of ascending order for first partition and descending order for second partition.

In Modified extended speech codec, the bit rate is actuated by avoiding two bits requirement of switching between forward mode and backward mode excitation codebook structure in standard (11.8 Kbps) speech codec. With the novel way of transmission of sign magnitude pulses of optimized excitation codevector in turn reduces the one-bit requirement in transmission of each of the two selected pulse positions indices in binary. Former standardised ITU-T CS-ACELP based speech codec requires three bits for transmission for each selected pulse positions as each track comprises of eight different pulse positions. The unique way of transmission of excitation codevector indices and sign magnitude pulse of the selected pulse positions proposes reduction of five bits per sub frame (40 samples) and ten bits per frame (80 samples) transmission in excitation codevector indices and actuates the bit rate of 10.6 kbps. Reduction in transmission of ten bits per frame creates a room for transmission of confidential/hidden data or a data as ownership proof with the help of steganography or watermarking techniques of information hiding in the speech signal. One of the emerging and popular techniques for information hiding in recent generation mobile communication is discussed further in next section.

7 Introduction to Quantization Index Modulation (QIM)

As quantization is a lossy process, the quantization noise which is introduced at the transmitter impinges an adverse effect on recovered quality of the speech signal when the speech is synthesized at the receiver end. Quality of the synthesized speech affected by the quantization noise is much more annoying for the human auditory system in comparison with the deterioration caused in the synthesized speech quality by the white random noise. To avoid the distortion

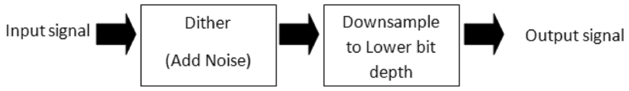

Fig. 2 Process of dithering

caused by the quantization noise, the random noise is added before reducing its bit depth; this process is called as dithering. Dithering is performed to make quantization noise more random or whiter. The basic idea of dithering is depicted in Fig. 2.

Audio recorded at 24 bits cannot be played because the CD Audio supports the 16 bits per sample format, hence to downsample and to convert it to a CD audio format, usually it is dithered with random noise when it is reduced from 24 to 16 bit so as to make the quantization error random in nature. It is also dependent on dynamic range of the material. One of the examples is a classical music which is having many quiet passages. Effect of Quantization noise is more in least significant bits. A 16 bit CD is having 65,536 levels of details, if quantization noise exists, it only going to be about 1 over 65,536 of maximum level of details. If it has to be noticed it should not be weaker than the other materials which are played. Due to this reason classical musical quiet passages are more susceptible to the quantization noise. On the other hand the quantization noise is drained out in today's latest electronic music, which has the super wall of guitar, super compress drums and a loud vocal which all are going at once. As quantization noise is very small it causes worse effects on quiet control noise free environment like silent passage of the speech signal. In proposed scheme of information hiding, quantization index modulation is applied on the saved 10 bits of excitation codevector which was discussed in Sect. 8.

8 Information hiding in line spectrum frequency (LSF) parameter of speech signal

Information hiding in the proposed CS-ACELP speech codec is achieved by determining one of the parameter of the speech signal which is parameterized at transmitter side and recovered at receiver side to synthesize the coded speech signal. Line prediction coefficients are one of the parameter of the speech signal which is derived from linear prediction analysis that is the most powerful speech analysis technique.

8.1 Introduction to line spectral frequency (LSF)

Linear Prediction (LP) has become the predominant technique (ITU-T Recommendation, 2007) for estimating the basic speech parameters like formants, pitch, spectra and vocal tract area functions; and also for representing a speech for low bit rate transmission or storage. LPC is also referred as an "inverse filtering" as its aim is to determine the "all zero" filter response which is the inverse of the vocal tract model. The hybrid speech coders which are working on the principle of analysis by synthesis (ABS), require to update the excitation optimization interval in the time slot of 4 to 7.5 ms which is in fact less than the LPC coefficients update frame size. To avoid this remedy, interpolation is performed to update the filter parameters on every subframe basis, while transmitting them to the decoder with a time slot of one frame.

As interpolation is performed on a subframe basis, the set of predictor coefficients cannot be used for interpolation because predictor coefficients are transmitted with a time slot of one frame. Interpolation of predictor coefficients may result into instability of a synthesis filter, which may cause a serious distortion at the time of the reconstruction of the speech signal. Interpolation is, therefore performed using a transform set of parameters where the filter stability can be easily guaranteed by using the Linear Spectral Frequency.

8.2 Technique of information hiding in LSF parameter of the speech signal using DM-QIM

8.2.1 Information hiding in quantized linear spectrum pair coefficients

In legacy speech codec, quantized and un-quantized Linear Prediction (LP) coefficients are parameterized from the original speech signal for interpolation and updating purposes of buffer for synthesizing the speech parameters at the weighted synthesis filter stage of encoder to be made compatible in operation with analysis by synthesis principle. Among the two forms of LP coefficients, un-quantized LP coefficients are targeted and utilized for achieving information hiding in speech signal. Introducing information hiding in quantized LP coefficients may lead to mismatch of the aimed secret data at the receiver and also cause the distortion in the output coded speech signal as quantized LP coefficients undergo the block quantization (Vector Quantization) for determining the codeword for line spectrum pairs named as L0, L1, L2 and L3. At transmitter side the un-quantized LSP coefficients are represented in LSF representation within the normalized frequency range of 0 to π by ITU-T Recommendation (2007);

$$\omega_i = \arccos\left(q_i\right) \quad i = 1, \ldots, 10 \tag{6}$$

Technique for information hiding in un-quantized LSP coefficients is discussed in Sect. 8.2.1.1.

8.2.1.1 Information hiding using dither modulated-quantization Index Modulation (DM-QIM) The un-quantized LSF coefficients are dithered by introduction of two different step sizes classified as Δ/4 and −Δ/4. This per frame coefficient dithering is performed to hide in turn to indicate bit ‘1’ or bit ‘0’, which solely dependent on an application or requirement of the user. The conventional form of QIM is given by Wang and Unoki (2013);

$$D(s) = \Delta\left[\frac{s}{\Delta}\right] \tag{7}$$

where s is the input signal and Δ is the quantization step size, and “[]” indicates the value inside the symbolic representation is rounded with nearest value. Two different dither quantizers (Eqs. 8 and 9) are applied on the input speech signal for indication of hidden bit ‘0’ or ‘1’ (Wang & Unoki, 2013).

$$D_0(s) = D\left(s - Q_0\right) + Q_0 \tag{8}$$

$$D_1(s) = D\left(s - Q_1\right) + Q_1 \tag{9}$$

As 10 LSF coefficients are there per frame time slot of 10 ms, 10 bits information can be incorporated per frame transmission and reception. among these dithered 10 coefficients which actually indicates the 10 different bits in terms of ‘0’ and ‘1’, 8 coefficients are dithered to transmit the actual confidential data like some predetermined pattern for authorization of data, identification number for some secret quantity or tag, which is nothing but the information hiding in terms of steganography or watermarking. While remaining 3 coefficients are dithered to help receiver to understand the pattern of decoding the secret data with indication of application of information hiding, which is discussed in Sect. 8.2.1.2.

8.2.1.2 Method of transmission of secret data With the creation of 10 bits span in proposed CS-ACELP speech codec, the coefficients are dither modulated by the bits of ‘0’ or ‘1’ (Eqs. 8 and 9). Among the 10 bits, 3 bits are reserved and in a way dithered to provide indication regarding application of information hiding that transmitter wants to convey to the receiver and to provide how that secret data can be decoded correctly without any mismatch for interpretation.

The remaining 7 bits are transmitted with different combination to avoid the suspicion about the secret data to eavesdropper. From the array of 10 of Line Spectrum Frequency, the MSB coefficient and the last two coefficients are selected for dithering to represent the different combinations of applications of information hiding which could be interpreted in a special way according to the combination listed in Table 6.

The un-quantized LSF coefficients are dithered according to the transmitter requirement by following the notation of combination shown in Table 6.

For example, if transmitter wants to share the steganography secret id “337” with the receiver. The above id can be shared with receiver using 2,2,3 pattern of transmission form Table 6. As per the Table 6, to transmit the above confidential id, the pattern required to transmit to the receiver would be “0111111100”. The un-quantized LSF coefficients of that particular frame are dithered (using Eqs. 8 and 9) according to the pattern to be transmitted to the receiver. These dithered LSF coefficients are converted back to un-quantized LSP coefficients for transmission to the receiver.

8.3 Blind detection of steganography data or ownership proof pattern at receiver

As the information hiding phenomenon is incorporated in the speech codec, proposed scheme deals with the limited number of bits for information hiding. On the same note, the base codec legacy CS-ACELP 8 kbps speech codec which is used as platform for building up a proposed 10.6 kbps CS-ACELP speech codec is a popular and famous speech codec for voice over internet protocol (VoIP). As today’s world is a world of internet, transmission of secret data through VoIP

Table 6 Combination of selecting application of information hiding with the novel way of transmission of secret data

Application of information hiding		Different patterns of transmission of secret data (7 bits)	Combination of last two LSB’s	Final 3-bit combination
Steganography (MSB = ‘0’)	Watermarking (MSB = ‘1’)			
0		2,2,3	00	000
1		2,2,3	00	100
0		2,3,2	01	001
1		2,3,2	01	101
0		3,4	10	010
1		3,4	10	110
0		4,3	11	011
1		3,4	11	111

may get hacked or destroyed by the attacker. With limitation on transmission of number of confidential data bits through VoIP puts restriction on sharing the logic of information hiding between transmitter and receiver. To transmit secret data without any mismatch to the receiver and also without compromising the coded speech quality, non-blind detection logic of secret data is incorporated at the receiver.

In blind detection process, only knowledge regarding the combination of selection of application of information hiding and different combinations depict different pattern which is shown in Table 6 is known to transmitter and receiver globally. At the receiver, firstly received bit stream of LSP parameter is decoded and un-quantized and quantized LSP coefficients are retrieved back. Retrieved un-quantized dithered LSP coefficients are converted to un-quantized dithered LSF coefficients using Eq. (6). DM-QIM is applied once again on each un-quantized LSF coefficient with both the dither vectors of ‘1’ and ‘0’. After applying DM-QIM on each LSF coefficient with both the dither vectors the Euclidian distance is measured between each of dithered LSF coefficient with respect to decoded un-quantized LSF coefficients. The smallest distance among the two-calculated distance for dither vector ‘0’ and ‘1’ with respect to obtained un-quantized LSF coefficients indicates the embedded bit in each LSF coefficients. With the detection of embedded bit in each LSF coefficient determines and retrieves the 10-bit secret binary pattern which was actually transmitted by transmitter. The retrieved binary pattern is interpreted by the receiver with globally shared knowledge of Table 6.

9 Subjective and objective measures

Subjective and objective quality assessment measures are utilized for evaluation of overall performance of proposed codec incorporating information hiding. Subjective analysis is checked by MOS while objective estimation is classified into perceptual based, waveform based and spectral based analysis (Tahilramani & Bhatt, 2017).

9.1 Subjective measures

In subjective measure, MOS (Mean Opinion Score) is employed to determine the quality of output codec speech. The quality of the output speech is asked to judge by 30 different listeners (out of which 15 are male and 15 are female listeners). The subjects are requested to rate the overall eminence of the speech (as shown in Table 7) by listening the speech in silent environment with superior quality of head phones (Tahilramani & Bhatt, 2017).

Table 7 Mean opinion score (MOS) ratings (Tahilramani & Bhatt, 2017)

S. no.	Choice	MOS ratings
1	Excellent	5
2	GOOD	4
3	Fair	3
4	Poor	2
5	Bad	1

9.2 Objective measures

Performance of proposed speech codec with the provision of confidential information hiding is evaluated with different objective quality assessment measures classified into waveform, spectral, perceptual and composite measures based analysis (Bhatt & Kosta, 2011; Tahilramani & Bhatt, 2015, 2017). In waveform based analysis: Mean square Error (MSE), Segmental SNR (SNRSEG), Absolute Error (ABS), Signal to Noise Ratio (SNR), Root Mean Square Error (RMSE) (Tahilramani & Bhatt, 2017); while in Perceptual based analysis: Perceptual Evaluation of Speech Quality (PESQ) (ITU-T Recommendation, 2000; Tahilramani & Bhatt, 2017) and in Spectral based analysis Log Likelihood Ratio (LLR), Itakura Saito Distance (ISD), Cepstrum Distance (CEP), Frequency Weighted Segmental SNR (fwSNRseg), Weighted slop spectrum distance (fwSNRs) (Bhatt & Kosta, 2012; Tahilramani & Bhatt, 2017) are explored. In Composite Measures parameters like Csig, Cbak, Covl are computed as per (Tahilramani & Bhatt, 2017).

10 Simulation of proposed coder

Here, proposed 10.6 Kbps speech codec with and without information hiding is realized in MATLAB and performance of proposed codec with comparative analysis is evaluated using different subjective and objective measures. For the sake of analysis of subjective and objective parameter, five different wave files from the Voxforge speech corpus database have been chosen (http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Original/16kHz_16bit/). Each sample of a wave file is represented by 16 bits with sampling rate of 16 KHz. The proposed CS-ACELP speech codec creates the room of 10 bits/s for information hiding shown in Sect. 10.3.

10.1 Result obtained for MOS analysis

Subjective analysis is performed for twenty different wave files taken from Voxforge speech corpus database (http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Original/16kHz_16bit/) and it is shown in Fig. 3. From the Fig. 3, It can be observed that, the MOS scores for all decoded output speech of proposed speech codec with and without information hiding are quite

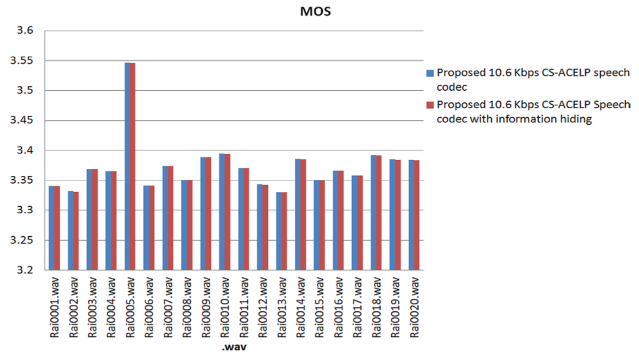

Fig. 3 MOS score comparison between proposed 10.6 kbps CS-ACELP Speech codec and proposed 10.6 Kbps with information hiding

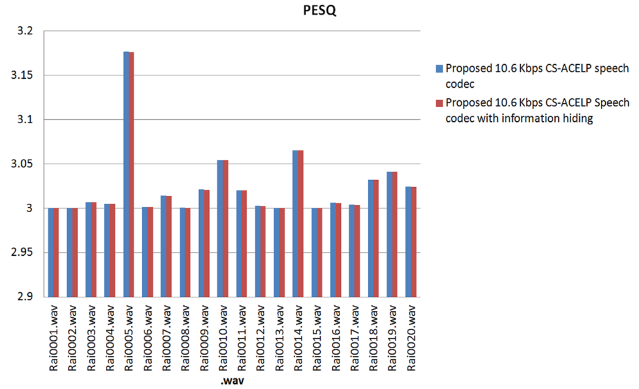

Fig. 4 PESQ score comparison between proposed 10.6 Kbps CS-ACELP Speech codec and proposed 10.6 Kbps with information hiding

satisfactory. In addition, it is observed that the Inclusion of information hiding using Quantization Index Modulation in per frame of speech parameterization does not affect the quality of output decoded speech signal in a big way.

10.2 Result obtained for objective analysis

The results of the PESQ (perceptual evaluation of speech quality) for the proposed speech codec (10.6 Kbps) with and without information hiding are shown in the Fig. 4, which is classified as one of the optimum parameter for objective quality assessment parameter of the speech signal. From the results it can be witnessed that the results of the PESQ of proposed speech codec with and without information hiding are fair enough to be justified with the results of the subjective analysis results of MOS. The results of the categorization of objective quality assessment parameters are shown in Table 8, Table 9 and Table 10. The outcomes of the classification of different objective quality assessment parameters are also precisely good, accurate and comparable for proposed speech codec for without and with embedding confidential data.

10.3 Secret data embedding capacity in proposed CS-ACELP speech codec

As depicted in Sect. 4, codebook partition and label assignment approach used for transmission of excitation codevector indices and the sign of the respective excitation codevector creates a scope of 10 bits per second for information hiding in speech signal. As discussed in Sect. 8.2.1.2, 3 bits are required to transmit for the indication of the application of steganography or watermarking, while remaining 7 bits span is kept vacant for secret data transmission to the receiver.

Total 10 unquantized LSF (Line Spectral frequency) coefficients are transmitted per frame. As described in Sect. 8.2, these LSF coefficients are dither modulated using Quantization index modulation with the fixed step size of Δ/4 and −Δ/4. As total 10 are transmitted per frame time slot of 10 ms, the information hiding capacity turn out be as 1 Kbps for steganography or watermarking data transmission.

Table 8 Waveform based analysis

Algorithm	Wav file	No of samples	ABS	MSE	RMSE	SNR	SNRseg
Proposed 10.6 Kbps CS-ACELP	Rai0007.wav	75,840	36.8241	0.000300	0.0170	4.5478	2.875424
	Rai0008.wav	82,560	46.2468	0.000548	0.0260	4.7112	2.501456
	Rai0009.wav	88,640	71.8554	0.000195	0.0145	4.4175	2.654712
	Rai0010.wav	73,440	48.2190	0.000730	0.0289	4.9959	3.702158
	Rai0011.wav	114,079	31.1123	0.000498	0.0240	4.3847	3.694123
Proposed 10.6 Kbps CS-ACELP with information hiding	Rai0007.wav	75,840	35.4380	0.000543	0.0233	3.0150	2.11844
	Rai0008.wav	82,560	55.9225	0.000110	0.0104	3.4914	2.44045
	Rai0009.wav	88,640	75.5609	0.000364	0.0191	3.2371	2.29741
	Rai0010.wav	73,440	51.5861	0.000150	0.0122	3.1149	2.19261
	Rai0011.wav	114,079	30.7245	0.000100	0.0100	3.0367	2.13972

Table 9 Perceptual based analysis

Algorithm	Wav file	No of samples	Covl	Csig	Cbak	PESQ	MOS
Proposed 10.6 Kbps CS-ACELP	Rai0007.wav	75,840	3.6426	4.2517	2.9813	3.3741	3.7035
	Rai0008.wav	82,560	3.6259	4.2318	2.9214	3.3508	3.6908
	Rai0009.wav	88,640	3.6839	4.2641	2.9984	3.3887	3.7187
	Rai0010.wav	73,440	3.7589	4.2943	3.0954	3.3941	3.7340
	Rai0011.wav	114,079	3.6319	4.2497	2.9649	3.3702	3.7002
Proposed 10.6 Kbps CS-ACELP with information hiding	Rai0007.wav	75,840	3.6362	4.2413	2.9755	3.3738	3.6969
	Rai0008.wav	82,560	3.6187	4.2239	2.9188	3.3504	3.6854
	Rai0009.wav	88,640	3.6769	4.2568	2.9885	3.3884	3.7021
	Rai0010.wav	73,440	3.7416	4.2877	3.0881	3.3938	3.7265
	Rai0011.wav	114,079	3.6249	4.2341	2.9592	3.3699	3.6951

Table 10 Perceptual based analysis

Algorithm	Wav file	No of samples	LLR	WSS	fwSNRseg	ISD	CEP
Proposed 10Kbps CS-ACELP	Rai0007.wav	75,840	0.392856	27.587410	8.9854	0.3398 3	4.0112
	Rai0008.wav	82,560	0.426541	31.457412	8.6247	0.3054 5	4.1241
	Rai0009.wav	88,640	0.362252	29.650014	8.2014	0.2564 2	3.9845
	Rai0010.wav	73,440	0.412410	27.545412	8.0122	0.5412 1	4.8514
	Rai0011.wav	114,079	0.393227	30.289741	8.4813	0.3025 4	4.0005
Proposed 10.6 Kbps CS-ACELP with information hiding	Rai0007.wav	75,840	0.827412	54.254124	8.5412	0.6615	4.1945
	Rai0008.wav	82,560	0.886545	58.987451	8.5247	0.6541	4.1612
	Rai0009.wav	88,640	0.814568	55.120142	8.3177	0.8412	4.1210
	Rai0010.wav	73,440	0.981816	58.238542	8.0945	0.8745	4.5412
	Rai0011.wav	114,079	0.940955	59.202356	8.1454	0.1245	4.2124

Information hiding of proposed speech codec is fixed as number of LSF coefficients transmitted are also fixed.

10.4 Introduction to calculation of population mean for objective and subjective result analysis of proposed speech codec

Consistency evaluation of proposed speech codec without information hiding is calculated using 99% confidence interval by incorporating the results of the subjective and objective parameters in the mathematical calculation. Results of PESQ (objective analysis) and MOS (subjective analysis) are considered as a test samples as an input for calculation of population mean. To calculate the population, mean based on confidence interval, different statistical parameters are calculated (Morgan et al., 2017) initially in order to calculate the range of the confidence interval. In order to calculate the population, mean of sample size less than 20, the t distribution table is followed in statistical mathematical calculation.

Population mean is derived mathematically as (Morgan et al., 2017),

Population mean = sample mean ± sample error

$$\mu = \ddot{x} \pm \frac{t.s_x}{\sqrt{n}} \tag{10}$$

where $\ddot{x}$ is a sample mean and μ is a population mean for calculating the range of 99% confidence interval. Critical value of t is obtained from t distribution table (Morgan et al., 2017). n is termed as sample size and s_x is called as standard deviation.

From Eq. (10), the standard deviation is calculated as,

$$s_x = \sqrt{\frac{\sum (x_i - \ddot{x})^2}{n - 1}} \tag{11}$$

To find out the value of t from t distribution table (Morgan et al., 2017), the degree of freedom (Morgan et al., 2017) is considered which is calculated as n – 1. As 5 PESQ and MOS results are calculated for 5 different wave files, the sample size for calculation of population mean is 5 (n) and degree of freedom is identified as 4 (n – 1). By considering the value of degree of freedom as 4 for calculation of 99% confidence interval the value of t from the t distribution table is determined as 4.596.

10.4.1 Calculation of population mean for objective quality assessment parameter PESQ for prosed 10.6 KBPS CS-ACLEP speech codec

From Eq. (11) the value of the standard deviation results in to 0.017.

From the Eq. (10), the range of the 99% confidence interval for PESQ samples of proposed speech codec without information hiding is (3.341, 3.411), which actually incorporates all the sample values of objective quality assessment parameter PESQ termed as input samples x_i.

10.4.2 Calculation of population mean for subjective quality assessment parameter MOS

From Eq. (11) the value of the standard deviation results in to 0.017.

Calculation of standard deviation for 5 PESQ samples as a test samples in population mean for 99% confidence interval is shown in Table 10.

The range for the 99% confidence interval for MOS samples of proposed speech codec without information hiding is (3.674, 3.744), which covers all the sample values of objective quality assessment parameter MOS termed as input samples x_i.

Here calculation of population mean for 99% confidence interval is shown for proposed speech codec without information hiding by considering results of PESQ and MOS as test samples as an input to the statistical system calculation. But the range of PESQ and MOS result samples considering proposed speech codec with information hiding also ensures consistency with population mean for 99% confidence interval.

11 Discussion and concluding remarks

CS-ACELP is analysis by synthesis based low bit rate speech codec which is very much popular nowadays in VoIP application used for secure speech communication between transmitter and receiver for the provision of voice plus data traffic over a network.

In proposed work, modification in existing excitation codebook structure of legacy speech codec is proposed for the purpose of reducing the rigorous searching complexity with more numbers of searches required in legacy speech codec due to the more number of pulses in final track. The least significant pulse replacement approach is used for reducing the number of searches required form 320 in exhaustive focused search approach of legacy coder to 60 number of searches in proposed codec excitation codebook structure. The unique approach of codebook partition and label assignment is applied for the transmission of indices of the optimized excitation codevector with its respective sign which in turn reduces the number of bit require to transmit the indices of excitation codevector. The novel approach of transmission of indices of excitation codevector creates the room of 10 bits per frame for steganography or ownership data transmission.

The proposed speech codec is clubbed with the very famous and new generation data hiding technique called as Quantization Index Modulation to incorporate data hiding in line spectral feature of the speech signal. In fact QIM technique makes the signal more whiter to avoid the distortion caused due to Quantization noise which is unavoidable lossy noise.

Here the results of the subjective and objective parameters are evaluated for proposed speech codec with and without information hiding. The results of the different subjective and objective analysis of two different proposed scheme are satisfactorily fair and the difference between the results of the two-proposed scheme is quite negligible. The accuracy and performance consistency of the proposed speech codec is measured with population mean for 99% confidence interval. The consistency is calculated in terms of range of confidence interval for PESQ and MOS results as a test samples. Results for the range of 99% confidence interval ensures the inclusion of all test samples in the range of population mean.

References

Bernard, A. P. (2005). Algebraic codebook system and method. *U.S. Patent US6847929B2*, January 25, 2005.

Bhatt, N., & Kosta, Y. (2011). Proposed modifications in ETSI GSM 06.10 Full Rate speech codec and its overall evaluation of performance using MATLAB. *International Journal of Speech Technology, 14*(3), 157.

Bhatt, N., & Kosta, Y. (2012). Implementation of variable bit rate data hiding techniques on standard and proposed GSM 06.10 full rate coder and its overall comparative evaluation of performance. *International Journal of Speech Technology, 16*(3), 285–293.

http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Original/16kHz_16bit/. Accessed 19 May 2021.

Hu, Y., & Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Speech and Audio Processing, 16*(1), 229–238.

ITU-T Recommendation. (2000). Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs (ITU-T Rec. P. 862, 2001).

ITU-T Recommendation. (2003). Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm (ITU-T Rec. P. 835, 2003).

ITU-T Recommendation. (2007). Coding of speech at 8 kbit/s using conjugate-structure algebraic-cod-excited linear prediction (CS-ACELP).

Lee, E.-D., & Kim, D.-Y. (2010). Method for searching fixed codebook based upon global pulse replacement. *U.S. Patent US8185385B2*, April 26, 2010.

Morgan, D., Permutt, T., & Seldrup, J. (2017). Pharmaceutical statistics. *The Journal of Applied Statistics in the Pharmaceutical Industry, 16*(3), 615.

Salami, R., Laflamme, C., Adoul, J.-P., Kataoka, A., Hayashi, S., Moriya, T., Lamblin, C., Massaloux, D., Proust, S., Kroon, P., & Shoham, Y. (1998). Design and description of CS-ACELP: A toll quality 8 kb/s speech coder. *IEEE Transaction on Speech and Audio Processing, 6*(2), 116–130.

Tahilramani, N., & Bhatt, N. (2015). Steganography in speech signal with enhanced multi-pulse excitation code vector with reduced number of bits. In *IEEE international conference on electrical, electronics, signals, communication and optimization* (*EESCO*) (pp. 1–4, 24–25).

Tahilramani, N. V., & Bhatt, N. (2017). Proposed modificationsin ITU-T G.729 8 kbps CS-ACELP speech codec and its overall performance analysis. *International Journal of Speech Technology*. https://doi.org/10.1007/s10772-017-9431-3

Wang, S., & Unoki, M. (2013).Watermarking method for speech signals based on modi cations to LSFs. In *Proc. IIHMSP* (pp. 283–286).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.