CrossMark

Proposed modifications in ITU-T G.729 8 kbps CS-ACELP speech codec and its overall performance analysis

Nikunj Tahilramani[1,2] · Ninad Bhatt[3]

Abstract This paper proposes modification in the transmission of excitation codevector and its non-zero pulse sign magnitude using "codebook partition and label assignment" approach, which in turn reduces the number of bits required to transmit it through the communication channel in legacy CS-ACELP 8 kbps speech codec. The proposed approach uses the excitation codebook structure of forward mode standard G.729E 11.8 kbps with two non-zero pulses per track which avoids the use of two algebraic codebook structure for forward mode as well as for backward mode of G.729E with least significant pulse replacement approach for finding optimized excitation codevector. Proposed modification in legacy 8 kbps CS-ACELP (80 bits/10 ms) speech codec actuates the bit rate of 10.6 kbps (106 bits/10 ms) with a better objective and subjective analysis in stark contrast with legacy 8 kbps CS-ACELP speech coder and also avoids the switching of codebook modes of standard 11.8 kbps (G.729E) CS-ACELP speech coder. This paper also aims to propose the reduction in the number of searches in the final codevector of excitation structure by considering initial codevector as a final codevector which improves the quality of the speech compared to the output speech quality of legacy G.729 CS-ACELP working at 8 kbps. Both legacy CS-ACELP 8 kbps speech codec and proposed CS-ACELP 10.6 kbps are implemented in MATLAB. Subjective and objective analysis are carried out on a proposed CS-ACELP 10.6 kbps speech codec in order to evaluate its performance and the results obtained are then cross-compared with the results of legacy CS-ACELP (8 kbps) using set of tables and graphs. It is evident from obtained results that both PESQ and MOS scores are quite comparable for each set of wave files even though bitrates are reduced. Consistency and efficiency of proposed algorithm is assured by calculating the population mean of 95% confidence interval based on obtained objective and subjective parameter results.

Keywords Excitation codevector · CS-ACELP · ITU-T · MATLAB · Subjective analysis · Objective analysis · Population mean · Confidence interval

1 Introduction

Conjugate structure algebraic-code-excited linear prediction (CS-ACELP) speech coder basically categorized in to Hybrid coder (Wu and Yang 2006; analysis by synthesis coder) classification which provides attractive trade-off between waveform coders and vocoders with satisfactory speech quality and transmission bit rate. Research and development in the domain of source coding techniques of CELP (code excited linear prediction), ACELP (algebraic CELP) and CS-ACELP (conjugate structure-ACELP) still continue to emerge as a popular area of research worldwide among researchers and academicians.

The speech encoder takes its input as a 16-bit linear PCM from the audio part of the mobile station or on the network side. The 64 k bit/s data, should be converted to

✉ Nikunj Tahilramani
   nikunjvtec@gmail.com

   Ninad Bhatt
   bhattninad@gmail.com

[1] Uka Tarsadia University, Bardoli, Gujarat, India

[2] Silver Oak College of Engineering and Technology, Ahmedabad, India

[3] Department of Electronics and Communication, C. K. Pithawalla College of Engineering & Technology, Dumas Road, Surat, Gujarat, India

Fig. 1 Detailed flow of excitation sequence of CS-ACELP based speech coder
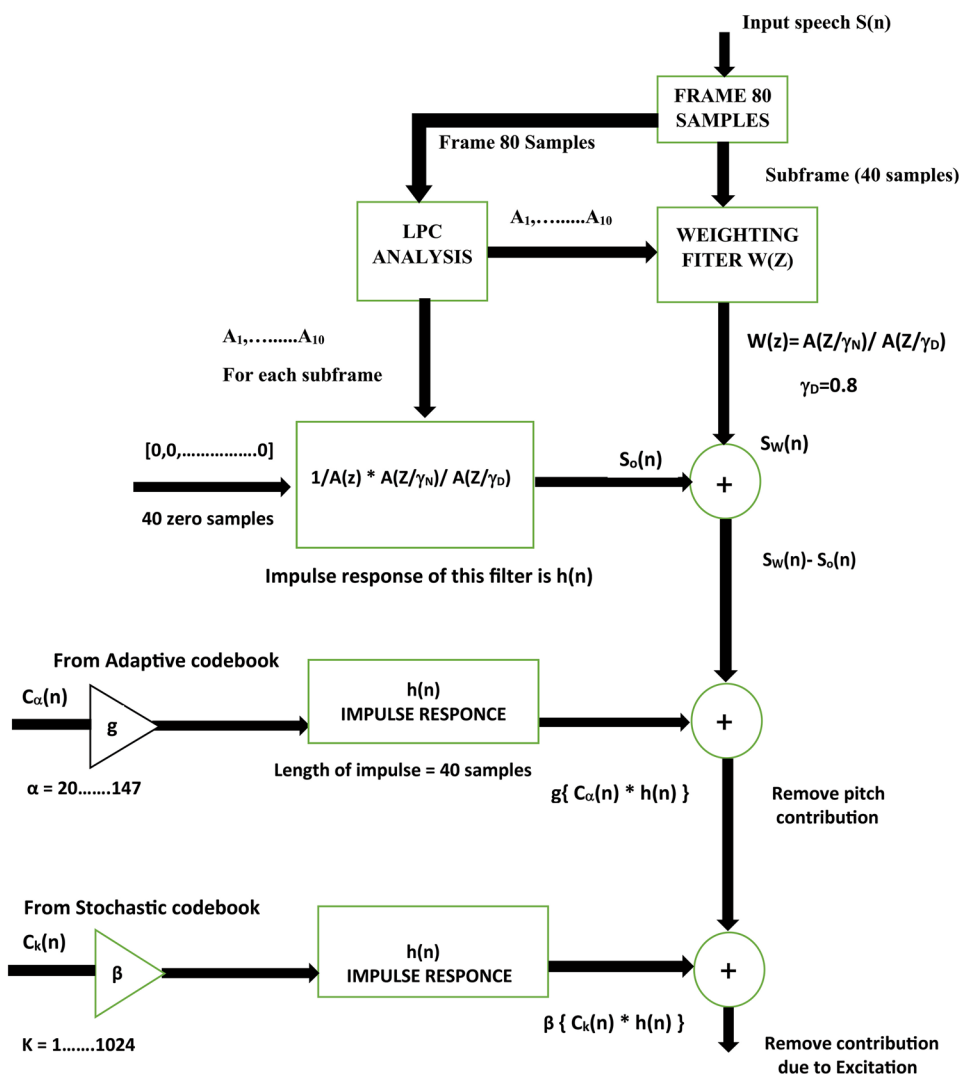
16 bit linear PCM before encoding, and 16 bit linear PCM to the appropriate form after reconstruction. CS-ACELP describes the trailed mapping between input blocks of 160 past speech samples, 80 present speech samples and 40 future speech samples in 16 bit linear PCM format to encoded blocks of 80 bits to output blocks of 240 total speech samples (ITU-T Recommendation 2007). The rate of input sampling is 16,000 samples/s which is processed for a coding operation at 8000 samples/s and after reconstruction once again it is interpolated to 16,000 samples/s. The rate of sampling is 16,000 samples/s leading to an average bit rate of 8 kbps.

This paper is organized as follows. In Sect. 2, excitation sequence of CS-ACELP based speech codec is introduced. In Sect. 3, excitation codebook structure of legacy 8 and 11.8 kbps CS-ACELP based speech codec is touched upon. Section 4 highlights the proposed modification suggested in

the search engine to determine the optimized codevector. In Sect. 5, various subjective and objective quality assessment parameters are depicted. Section 6 deals with comparative performance evaluation of CS-ACELP speech coder which is computed and demonstrated using set of graphs and tables. Finally, the concluding remarks are given in Sect. 7.

2 CS-ACELP speech codec excitation sequence generation

The detailed flow of excitation sequence of CS-ACELP based speech codec is shown in the Fig. 1. The input speech signal frame, consisting of 80 speech samples is first pre-processed in which mainly signal scaling and high pass filtering operations are performed to reduce the overflows in fix point implementation and to remove the undesired

Table 1 Bit allocation of the 8 kbit/s CS-ACELP algorithm (10 ms frame; ITU-T Recommendation 2007)

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

Table 2 Fixed codebook excitation structure of legacy 8 kbps CS-ACELP based 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

low frequency components. LPC (linear predictive coding) analysis is performed over this 80 speech samples to determine ten LPC coefficients per frame which is a short-term analysis of the speech signal. The combination of per frame LPC coefficients and original segmented speech samples is passed through a weighting filter which take cares of the perception of the human auditory system. Generally, the noise located in a lower energy region is much more annoying when the speech is reconstructed with short term synthesis filter. The error weighting filter shifts these kinds of noise from lower energy region to higher energy region so that it is masked at the time of reconstruction of the speech signal. Weighted error output is subtracted from the output of impulse response of the filter which is a combination of short term synthesis filter and error weighting filter. Initially the impulse response is generated by 40 zero input samples to the above hybrid combination to flush everything out of the system. The pitch contribution is removed from the hybrid output by subtracting the combination of adaptive codebook and impulse response from it. Firstly, the optimum open loop pitch and its gain are determined from the adaptive codebook and the input is forwarded to impulse response to generate the combination of codevector from adaptive codebook and impulse response. The output long term residuals which are not having any pitch information is again used in subtraction operation with the combination of a codevector from stochastic codebook

Table 3 Fixed codebook excitation structure of standard 11.8 kbps CS-ACELP based Speech Coder in forward LP mode (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

and its sign magnitude to remove the contribution of excitation. The subtracted output which is also called as short term residual is again feedback as a weighted error input to model the excitation structure. The Bit allocation for the legacy CS-ACELP working at 8 kbps is shown in Table 1.

3 Excitation codebook structure of legacy 8 kbps and extended 11.8 CS-ACELP based speech codec

ITU-T G.729 CS-ACELP speech coder is having 40 pulse positions with 4 track excitation codebook structure as depicted in Table 2. Excitation codebook structure is having 8 positions in first three tracks and 16 positions in final track. In the conventional 8 kbps, fixed codebook search procedure, the final codevector is determined by 8192 number of searches in case of Full search approach (Salami et al. 1998),1440 number of searches in case of Focused search approach (12), 320 number of searches in case of Depth first search approach (Adoul and Laflamme 1997) with exhaustive recursive searches.

The numbers of searches are more because of final track is having 16 pulse positions, which increases the alternate searches of the final track eight positions every time when the search procedure starts with different tracks. As first three tracks are having eight positions, it requires 3 bits to code these positions and final track pulse position requires 4 bits as it is having 16 pulse positions. Along with the four pulse positions from four different tracks, it also requires to transmit the sign of the corresponding pulse positions. On the whole it requires 13 bits to transmit final best codevector and 4 sign bits for respective selected pulse position per subframe.

The excitation structure of G.729 (8 kbps) is replaced by forward mode excitation structure of extended CS-ACELP 11.8 kbps speech codec which follows the excitation structure of 5 tracks with each track is having two non-zero pulses with a sign magnitude of ± 1. The excitation structure of standard extended 11.8 kbps CS-ACELP speech coder is demonstrated in Table 3.

ITU-T standardized extended G.729E (11.8 kbps) fixed codebook excitation structure, which uses different bit codebook structure namely as forward mode excitation codebook (35 bits) structure and backward mode fixed codebook excitation code structure (ITU-T Recommendation 2007; 44 bits). In the proposed analysis, the 35 bits fixed excitation codebook structure is used at transmitter side as well as receiver side. The excitation codevector is determined over a subframe of 40 samples. The algebraic codebook search is performed by least significant pulse replacement procedure (Bernard 2005). Initial fixed codevector is determined by choosing a pulse position having a largest magnitude of correlation vector d(n) (Eq. 2). The exhaustive search procedure is started by arranging the coefficients of correlation vector d(n) according to the pulse positions of a conjugate excitation codebook structure highlighted in Table 3. First pulse position among the two from the respective track is determined by the maximum of the eight positions of that particular track and second pulse each track is determined by finding the Euclidian distance from first maximum of each respective track. The next stage codevector is determined by replacing each and every pulse position of the initial codevector with the other pulse position of respective track one at a time. The combination of the pulse position which maximizes the value of (Eq. 1) Q_k is declared as a next stage codevector. 12 Q_k values are computed per track. The final best codevector is determined by the maximum value of Q_k out of 60 Q_k values (Bernard 2005)

$$\max_k Q_k = \max_k \frac{C_k^2}{E_k} = \max_k \frac{(d^t c_k)^2}{c_k^t \emptyset_{c_k}}$$

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

here, M is a number of tracks in a subframe analysis.

A Kth codebook vector is described as C_k and t denotes a transposed matrix. d is called as correlation vector and matrix PHI are described as (ITU-T Recommendation 2007):

$$d(n) = \sum_{i=n}^{M-1} x_2(i)h(i-n), \quad i = 0, \ldots, M \quad (2)$$

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

From Eqs. 2 and 3 the total number of pulse positions in a sub-frame is M, a target signal for the fixed codebook searching is expressed as $x_2(n)$ and an impulse response of a linear predictive synthesizing filter is described as h(n) (ITU-T Recommendation 2007).

Also a numerator and a denominator of Eq. 1 are described as (ITU-T Recommendation 2007):

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

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

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

Final codevector is having best two pulse positions per track which maximizes the value of Eq. 1, which require 6 bits per track as each track is having total of eight pulse positions. Apart from the pulse position, among the two non-zero sign magnitude pulse having amplitudes ±1, only one of the sign of the non-zero pulse is transmitted and other sign is derived directly at the decoder from the other one which has been received at the receiver, which is being explain below.

Let the sign and the positions of the two non-zero pulses are s1, s2, p1 and p2. In all the iterations, the smallest selected position among the two is assigned as p1 and the other one as p2. If p1 ≤ p2, then s2 = s1 else s2 is different from s1. At the time of finding a final codeword if both the signs are equal then p1 is declared as smallest among the two position otherwise p1 is declared as largest position and p2 is declared as smallest position (ITU-T Recommendation 2007).

4 Proposed modification in the search engine of the excitation codebook structure for determining optimised codevector

The proposed modification uses the codebook partition approach for modification in the searching procedure of determining final codevector. In the proposed approach, firstly the excitation codebook is partitioned into two equal parts in the structure dimensions. The standardized structure dimension of the fixed excitation codebook is having five tracks with total of 40 positions. The proposed approach uses the advantage of number of positions in each track which is an even number. Partitioning of a codebook structure allows the coding of the individual position into total of four combinations instead of eight combinations. This modification requires total of 20 bits for final excitation codevector and 5 sign bits for transmission through a channel per subframe.

4.1 Efficient transmission of position and sign magnitude of non-zero pulses with reduced number of bits

In the proposed codebook partition approach the two partitions are assigned with the four-different combination of 00, 01, 10 and 11. From the least significant pulse replacement approach the two non-zero elements are obtained per track. As codebook, partitioning approach is used to provide the coding using 2-bit position instead of 3-bit position, the selected two non-zero pulses may be either from a first partition of the excitation codebook or from a second partition of the excitation codebook. The label of combination is assigned in ascending order as 00, 01, 10 and 11 in the partition 1 of the codebook, while in a partition 2, labels assignment is starting from 11, 10, 01 and 00.

As codebook is partitioned into two equal parts, there are total three possibilities where the non-zero pulses are located. First possibility corresponds that the two non-zero pulses are found in the partition 1, similarly the second possibility give rise to investigation of a two non-zero pulses in partition 2. While the third possibility state the combination of finding the two non-zero pulses in two different partitions. The two non-zero pulses having amplitude ± 1 is transmitted along with the positions in their respective track.

Let the two non-zero pulses are located at positions 0 and 5 belongs to partition 1 in track 1 having sign magnitude $+1$ and -1. According to label assignment these two pulses are having binary label of 00 and 01. The corresponding sign bits are transmitted in 2 bits with 1 and 0 for the above case. the sign bits for the all the five tracks are transmitted according to 3 described possibilities. if the two non-zero pulses are from the same partition then the sign bits are transmitted in a 2 in 1cell which is same with the case of possibility 2. If the non-zero pulses are from the different partition as of possibility 3 then the sign magnitude of the two non-zero pulses are transmitted with 1 in 1 cell.

At the receiver, initially the corresponding cell format is identified, and if the cell format is of 2 in 1 cell then it is identified that the two non-zero pulses are located either in partition 1 or in partition 2. To confirm the partition number, the respective binary labels are identified with their order. If the 4-bit binary format of the codeword is of ascending order, then sign magnitudes are placed at those assigned corresponding binary locations in partition 1. The contradictory prevails the same for the partition 2 in descending order. Non-zero pulses belonging to different track are identified at receiver using 1 in 1 cell combination and they are placed at their respective binary combination but in two different partition.With the proposed modification in transmission of position and

sign magnitude of excitation codevector reduces 10 bits per frame for the transmission of the index of the excitation codevector. As only one excitation codebook is used at transmitter as well as at receiver, no switching bit is required, which actually required in the case of standard 11.8 CS-ACELP based speech codec due to two different codebook structures are used at encoder and decoder. The bit allocation of proposed 10.6 Kbps CS-ACELP based speech codec is shown in Table 4.

4.2 Reduction in a number of searches of excitation codevector of legacy 8 kbps CS-ACELP coder

With the proposed modification in transmission of excitation codeword with its sign magnitude, the excitation search can be also restricted by finding the initial codevector only. In the proposed modification, Initial codevector is determined by the position of the first and second maxima of correlation vector d(n) (Eq. 2.) from each track. If the initial codevector made up of ten non-zero pulses, is considered as a final codevector then no searches are required which was initially reduced to 60 from 320 of focused search approach with modified search procedure of excitation codebook structure. It is being observed in Figs. 2, 3 that, if the initial codevector is processed as a final codevector then it yields good quality speech at the decoder with satisfactory subjective and objective parameter results compared to the legacy 8 kbps CS-ACELP speech coder. The results are compared with the 8 kbps CS-ACELP speech coder only as the numbers of searches are reduced to 0. With the proposed modification in transmission of sign magnitude and the position required 4 bits for position and 2 bits for sign for individual code word per track as there are 2 non-zero pulses per track. CS-ACELP 11.8 kbps recommendation

Table 4 Bit allocation of the proposed 10.6 kbit/s CS-ACELP algorithm (10 ms frame)

Parameter	Subframe 1	Subframe 2	Total bits/frame
LSP			18
Adaptive-codebook delay	8	5	13
Pitch-delay parity	1		1
Fixed-codebook index	20	20	40
Fixed-codebook sign	10	10	20
Codebook gains (stage 1)	3	3	6
Codebook gains (stage 2)	4	4	8
Total			106

requires 70 bits with 60 bits for codeword and 10 bits for sign transmission per frame while with proposed modification in transmission require 50 bits for final codevector 10 bits for sign magnitude transmission at the receiver. Proposed modification in CS-ACELP working at 8 kbps leads to the reduction of the bits required to transmit the excitation codevector compared to the extended CS-ACELP (11.8 kbps) and also it reduces the complexity of the two-different excitation codebook structure required at transmitter and at the receiver of CS-ACELP working at 11.8 kbps.

5 Subjective and objective measures

To evaluate the overall performance of proposed modification, objective and subjective quality assessment parameters are utilized. In subjective evaluation, MOS analysis is being explored whereas objective evaluation is categorized into waveform based, spectral based and perceptual based analysis.

5.1 Subjective measures

In subjective measure, MOS (mean opinion score) which is used to determine the quality of compressed speech at the output of the decoder. The quality of the output speech is asked to judge by randomly 30 persons. They are asked to rate the quality of the speech signal according to the options available in Table 4. The listeners are asked to judge the overall quality of the speech by playing the speech in noiseless environment with high quality head phones. The ratings are demonstrated in the Table 5.

5.2 Objective measures

To evaluate and compare the performance of proposed speech codec with legacy speech codec in terms of decoded speech quality and also comparison between decoded speech quality between the two codec's by considering initial codevector as a final codevector in proposed 10.6 kbps

Table 5 Mean opinion score (MOS) ratings (Ninad and Kosta 2012)

Sr. no.	Choice	MOS ratings
1	Excellent	5
2	Good	4
3	Fair	3
4	Poor	2
5	Bad	1

speech codec, different types of objective measures have been carried out. Objective measure is classified into waveform, spectral, perceptual and composite measures based analysis (Ninad and Kosta 2012).

5.2.1 Waveform based analysis

Following quality assessment parameters are evaluated in this category.

(1) Absolute error (ABS) is mathematically defined as,

$$Abserr = \sum |s_i - s_o| \qquad (6)$$

(2) Mean square error (MSE) is mathematically expressed as,

$$MSE = \left(\sum \frac{(S_i - S_o)^2}{N} \right) \qquad (7)$$

(3) Root mean square error (RMSE) is mathematically expressed as,

$$RMSE = \sqrt{\left(\sum \frac{(S_i - S_o)^2}{N} \right)} \qquad (8)$$

(4) Signal to noise ratio is mathematically given as,

$$SNR = 10 \log_{10} \frac{\sum |S_i|^2}{\sum |S_i - S_o|^2} \qquad (9)$$

where S_i = input signal, S_o = decoded output signal and N = total no. of frames.

(5) Segmental SNR is mathematically given as,

$$SNR_{SEG} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left[\frac{\sum_{n-m_{j-N+1}}^{m_j} s^2(n)}{\sum_{n-m_{j-N+1}}^{m_j} [s(n) - \hat{s}(n)]^2} \right] \qquad (10)$$

where s(n) = input signal, $\hat{s}$ (n) = decoded signal, N = segment length, M = no. of segments and m_j = end of the current segment.

5.2.2 Perceptual based analysis

(1) Perceptual evaluation of speech quality (PESQ) PESQ algorithm uses psychoacoustic and cognitive models by using a synchronization scheme, this algo-

rithm time aligns the original and degraded speech signals, as misalignment could result in a false quality score. To compute the speech-quality degradation represented by the disturbance metric between the psychophysical representations of the reference and degraded speech samples, the cognitive model performs complex non-linear calculations. PESQ is designed to analyze specific parameters of audio, including time warping, variable delays, transcoding, and noise. PESQ score is computed as a liner combination of the average disturbance value D_{ind} and the average asymmetrical disturbance value A_{ind} as follows (Ninad and Kosta 2012):

$$PESQ = a_0 + a_1 D_{ind} + a_2 A_{ind} \tag{11}$$

where a_0, a_1 and a_2 are calculated using Multiple linear regression analysis.

5.2.3 Spectral based analysis

Following parameters are categorized to perform spectral based analysis (Ninad and Kosta 2012).

(1) Log likelihood ratio (LLR) is defined by following equation,

$$d_{LLR}(\overrightarrow{a_p}, \overrightarrow{a_c}) = \log\left(\frac{\overrightarrow{a_p} R_c \overrightarrow{a_p^T}}{\overrightarrow{a_c} R_c \overrightarrow{a_c^T}}\right) \tag{12}$$

where $\overrightarrow{a_c}$ is the LPC vector of the original speech signal frame and $\overrightarrow{a_p}$ is the LPC vector of the decoded speech signal frame, and R_c is the autocorrelation matrix of the original speech signal.

(2) Itukara Saito distance measure is mathematically defined as,

$$d_{IS}(\overrightarrow{a_p}, \overrightarrow{a_c}) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\overrightarrow{a_p} R_c \overrightarrow{a_p^T}}{\overrightarrow{a_c} R_c \overrightarrow{a_c^T}}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \tag{13}$$

where σ_p^2 and σ_c^2 are LPC gains of original and decoded signals (Ninad and Kosta 2012). The range of the IS value is limited between 0 and 100.

(3) Cepstrum distance (CEP)

It provides an estimation of distance between two log spectra. The Cepstrum coefficients can be obtained with the recursion procedure of LPC coefficients as using the given expression:

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k} \tag{14}$$

where, p denotes the order of the LPC analysis. An objective measurement from the Cepstrum coefficient can be computed with following expression (Ninad and Kosta 2012):

$$d_{CEP}(\overrightarrow{c_c}, \overrightarrow{c_p}) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}[c_c(k) - c_p(k)]^2} \tag{15}$$

where $\overrightarrow{c_c}$ and $\overrightarrow{c_p}$ are the cepstrum coefficient vector of the original and recovered signal. The range of the limitation of the Cepstrum distance was limited between 0 and 10.

(4) Frequency weighted segmental SNR (fwSNRseg)

$$fwSNRseg = \frac{10}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) \log_{10}\frac{|X(j,m)|^2}{\left(|X(j,m)| - \left|\overrightarrow{\hat{X}}(j,m)\right|\right)^2}}{\sum_{j=1}^{K} W(j,m)} \tag{16}$$

where $W(j,m)$ is denoted as weight placed on the jth frequency band, k denotes the number of bands, M denotes the total number of frames in the signal, $|X(j,m)|$ is denoted as weighted original signal spectrum in the jth frequency band at the mth frame, while $\left|\hat{X}(j,m)\right|$ is denoted as weighted decoded signal spectrum in the same band.

(5) Weighted slop spectrum distance is defined as,

$$fwSNRs = \frac{10}{M}\sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) \log_{10}(|s_c(j,m)| - |s_p(j,m)|)^2}{\sum_{j=1}^{K} W(j,m)} \tag{17}$$

In each frequency band weighted slop spectrum distance calculates the weighted difference between the spectral slops. Spectral slope is calculated as the difference between adjacent spectral magnitudes in decibels. $s_c(j,m)$ and $s_p(j,m)$ are denoted as spectral slope of jth frequency band at frame m of the original and decoded speech signal with total of 25 number of bands (Ninad and Kosta 2012).

5.2.4 Composite measures

Unlike the simple objective measures parameters, there are certain parameters which combine all objective measures to form a new measure called as composite measure. Composite measure is the linear combination of existing

objective measures to form a new objective measure which utilizes linear regression analysis. Following parameters are utilized and checked for the effective composite measure: a measure called as C_{sig} for signal distortion which is a linear combination of PESQ, LLR and WSS measures, a measure which is known as C_{bak} for background noise distortion which is a linear combination of PESQ, segSNR and WSS measures, a measure which is responsible for overall speech quality measurement called as C_{ovl} formed by linearly combining WSS, LLR and PESQ measures.

The multiple linear regression analysis of above three composite measure is shown below (ITU-T Recommendation 2003; Falk and Chan 2006; Salmela and Mattila 2004; Grundlehner et al. 2005).

$$C_{sig} = 3.903 - 1.029 \cdot LLR + 0.603 \cdot PESQ - 0.009 \cdot WSS \tag{18}$$

$$C_{bak} = 1.634 + 0.478 \cdot PESQ - 0.007 \cdot WSS + 0.063 \cdot segSNR \tag{19}$$

$$C_{ovl} = 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - 0.007 \cdot WSS \tag{20}$$

6 Simulation of proposed algorithm based on MATLAB

Here, both legacy CS-CELP working at 8 kbps and proposed CS-ACELP working at 10.6 kbps are implemented in MATLAB and performance of both coders is evaluated using different subjective and objective measures. Excitation structure of legacy CS-ACELP working at 8 kbps having four tracks is replaced with 5 track structure of excitation codebbok with two non-zero pulses in each track instead of 1 non-zero pulse in each track. The proposed 10.6 kbps CS-ACELP creates the room of 12 bits/frame in 118 bits of CS-ACELP 11.8 kbps for steganographic data transmission or better error concealment at channel coding level. For the sake of analysis of subjective and objective parameter, five different wave files have been chosen.[1] Each wav file is sampled at 16 kHz and coded by 16 bits per sample.

6.1 Result obtained for MOS analysis

MOS analysis is performed for twenty different wave files. Wave files are taken from VoxForge speech corpus database.[2] Thirty random subjects had to judge the quality of speech in noise free environment using very high fidelity headphones. The listeners had to give a score to all the decoded speech files which are CS-ACELP 8 kbps decoded speech as well as proposed CS-ACELP 10.6 kbps decoded speech. The results of the MOS score of twenty different wav files[3] are shown in a Fig. 2. As it can be observed from Fig. 2 that MOS score of proposed CS-ACELP 10.6 kbps speech coder is far better than the legacy CS-ACELP 8 kbps speech coder. The result of decoded speech quality of output speech considering initial codevector as a final codevector with no search in the search engine of excitation codebook is also shown (Fig. 2) and compared with proposed 10.6 kbps CS-ACELP speech coder with 60 searches and with legacy CS-ACELP 8 kbps speech coder which require different number of searches in different search approach. It is witnessed that the quality of decoded speech by considering initial codevector as a final codevector is better compared to the other different search methods of legacy 8 kbps speech coder but less than the proposed approach with 60 numbers of searches.

6.2 Result obtained for objective analysis

Perceptual evaluation of speech quality (PESQ) based objective analysis is performed on 20 different wave files taken from VoxForge speech corpus database.[4] The results of the different objective classified quality assessment parameter based on waveform based analysis; perceptual based analysis and spectral based analysis are highlighted in Tables 6, 7 and 8 for seven different wave files taken from VoxForge speech corpus database.[5] The results for the other classified objective quality assessment parameters analysis have been shown in Table 6 for legacy CS-ACELP 8 kbps speech coder, proposed CS-ACELP 10.6 kbps speech coder and proposed 10.6 kbps CS-ACELP based speech codec with consideration of initial codevector as a final codevector. It can be observed that there is a large amount of distortion when both the legacy CS-ACELP speech coder and proposed CS-ACELP speech coder are compared. The results of all quality assessment parameter are quite fair in case of proposed coder compared to legacy coder (Fig. 3).

6.3 Computing population mean for result analysis of subjective and objective parameters

Performance of proposed coder is evaluated using objective quality assessment parameter called as PESQ and

---
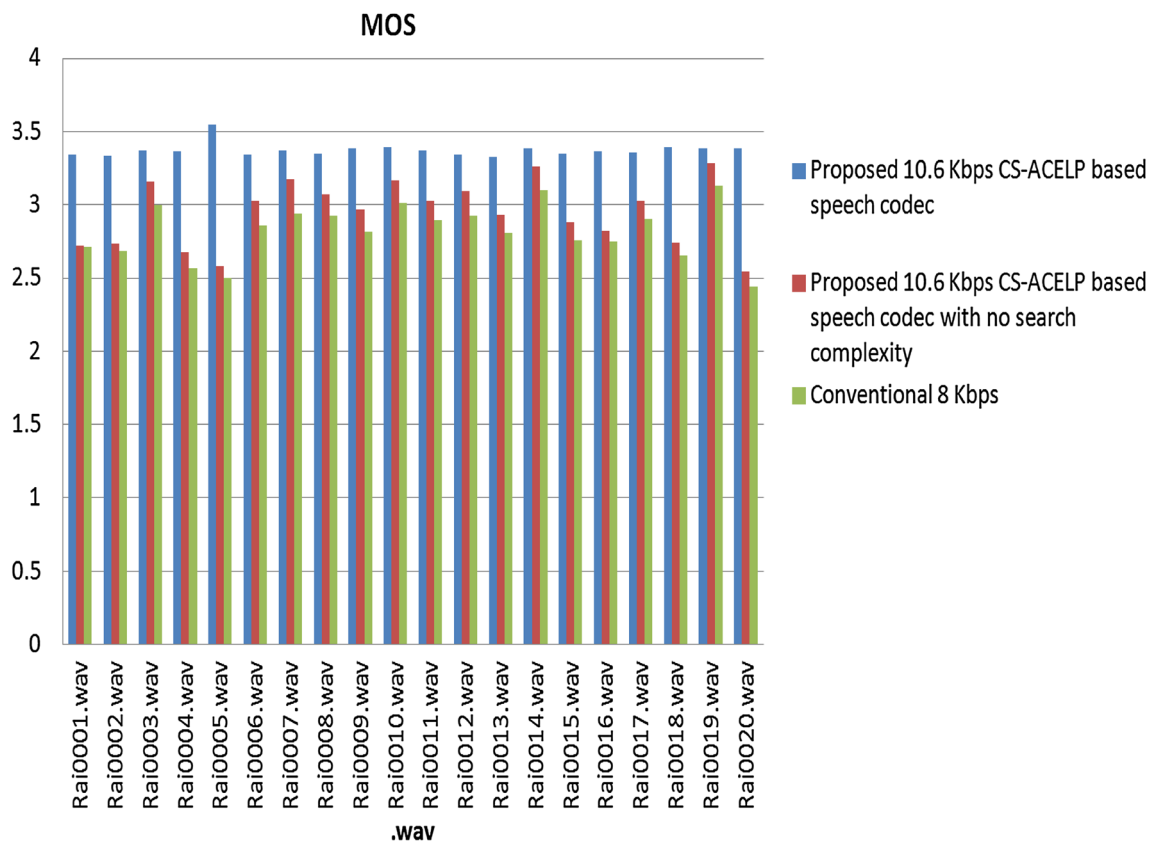
Fig. 2 MOS score comparison between proposed 10.6 kbps, proposed 10.6 kbps with no search complexity and legacy 8 kbps CS-ACELP based speech coder

subjective quality assessment parameter called as MOS. As can be witnessed from the Tables 6, 7 and 8 that the results of the different classified objective and subjective parameters are quite satisfactory. To evaluate performance oriented consistency of proposed coder, the population mean of 95% confidence interval is calculated based on the results obtained for subjective and objective parameters of seven different wave files in Sects. 2 and 3.

As demonstrated in (Morgan et al. 2017), range of confidence interval is calculated with the help of statistical parameters like mean, standard deviation, sample size and the standardized normal distribution values or critical t distribution values for different percentages of confidence interval. As per (Morgan et al. 2017), the sample size has direct relationship with the confidence interval. As depicted in (Morgan et al. 2017), when the sample size is small then in that case the critical values of t distribution are considered for confidence interval measurement.

Confidence interval is calculated as per (Morgan et al. 2017),

Population mean = sample mean $\pm$ sample error

$$\mu = \ddot{x} \pm \frac{t \cdot s_x}{\sqrt{n}} \tag{21}$$

where, μ is a population mean which defines the 95 or 99% confidence interval range. $\ddot{x}$ is a sample mean, t is decided from the critical values of t distribution table (Morgan et al. 2017), s_x is a standard deviation and n is defined as a sample size. The confidence interval calculation for 95 and 99% is shown in section A and section B.

6.3.1 Calculation of population mean of different confidence intervals for the proposed 10.6 kbps CS-ACELP based speech codec

6.3.1.1 Calculation of population mean based on objective speech quality evaluation parameter PESQ In below calculation following notations are followed:

x_i =7 PESQ values from Table 6 for proposed 10.6 kbps CS-ACELP based speech codec is used as seven samples which act as an input for population mean calculation.
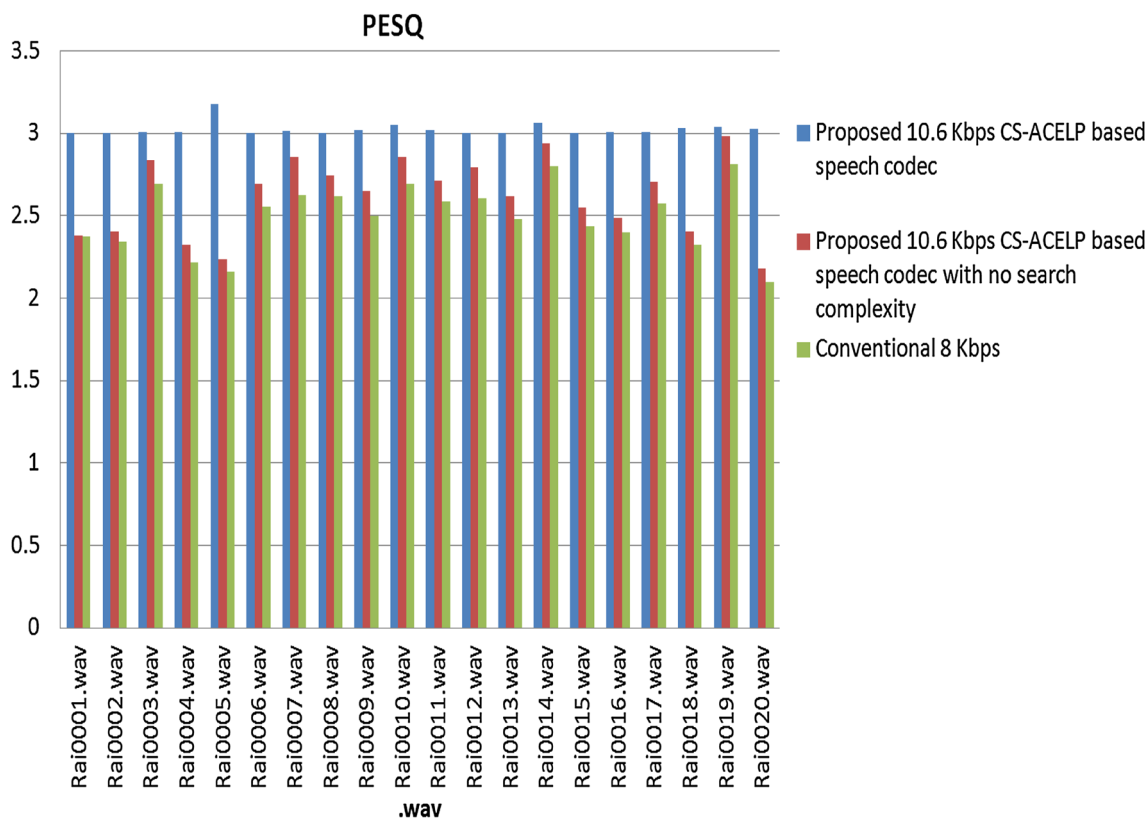
Fig. 3 PESQ score comparison between proposed 10.6 kbps, proposed 10.6 kbps with no search complexity and legacy 8 kbps CS-ACELP based speech coder

ẍ is sample mean,

n = sample size, which is 7 as per the total number of samples.

$$s_x = \text{standard deviation} = \sqrt{\frac{\sum (x_i - \ddot{x})^2}{n-1}} \qquad (22)$$

From the calculations the value of standard deviation (Eq. (22)) is $s_x = 0.064$.

(A) *Population mean for 95% confidence interval for PESQ samples*. Population mean is calculated from Eq. (21) in which value of t is taken as 2.447 by considering degree of freedom (Morgan et al. 2017) as 6 (n − 1) due to total of seven samples, from the table of critical values of t distribution (Morgan et al. 2017).

From Eq. (21) the value of population mean for 95% confidence interval turned out as (2.976, 3.094).

From the observation of the calculation of Table 9 and the calculation of confidence interval, it is observed that except 1 sample values (PESQ values), remaining six sample values resides in the range of population

mean of 95% confidence interval. The value of a 1 sample is 3.1768 which is a PESQ value of Rai0005.wav,[6] is beyond the range of highest value (3.094) of population mean of 95% confidence interval.

6.3.1.2 Calculation of population mean based on subjective speech quality evaluation parameter MOS   From the calculations the value of standard deviation [Eq. (22)] is $s_x = 0.073$.

(B) *Population mean for 95% confidence interval for MOS samples*. Population mean is calculated from Eq. (21) in which value of t is taken as 2.447 by considering degree of freedom as 6 (n − 1) (Morgan et al. 2017) due to total of seven samples, from the table of critical values of t distribution (Morgan et al. 2017).

From Eq. (21) the value of population mean for 95% confidence interval turned out as (3.321, 3.457).

From the observation of the calculation of Table 10 and the calculation of confidence interval, it is observed that except 1 sample values (MOS values), remaining six sample values resides in the range of population

---

6   See footnote 1.

Table 6 Waveform based analysis

Algorithm	Wav file	No. of samples	ABS	MSE	RMSE	SNR	SNRseg
Proposed 10.6 kbps CS-ACELP	Rai0004.wav	66,400	37.8321	0.000274	0.0166	4.4633	2.785259
	Rai0005.wav	89,920	56.2466	0.000647	0.0254	4.7048	2.480157
	Rai0016.wav	135,360	78.7124	0.000175	0.0132	4.3132	2.542308
	Rai0018.wav	95,040	48.2158	0.000765	0.0277	5.0441	3.585854
	Rai0020.wav	60,001	30.0050	0.000512	0.0226	4.4891	3.585854
	Rai0001.wav	120,640	70.2107	0.000946	0.0308	4.1778	3.014555
	Rai0002.wav	101,760	62.3476	0.000480	0.0219	4.8406	3.321254
Proposed 10.6 kbps CS-ACELP with no searching complexity	Rai0004.wav	66,400	38.4380	0.000543	0.0233	3.0150	2.11844
	Rai0005.wav	89,920	55.9225	0.000110	0.0104	3.4914	2.44045
	Rai0016.wav	135,360	75.5609	0.000364	0.0191	3.2371	2.29741
	Rai0018.wav	95,040	51.5861	0.000150	0.0122	3.1149	2.19261
	Rai0020.wav	60,001	30.7245	0.000100	0.0100	3.0367	2.13972
	Rai0001.wav	120,640	70.7233	0.000180	0.0420	3.2954	2.25413
	Rai0002.wav	101,760	60.6217	0.000906	0.0301	3.2547	2.24514
Legacy 8 kbps CS-ACELP	Rai0004.wav	66,400	38.4595	0.000520	0.0228	2.7852	1.718748
	Rai0005.wav	89,920	56.0021	0.000110	0.0105	2.3989	1.601923
	Rai0016.wav	135,360	75.4704	0.000337	0.0184	2.8989	1.689363
	Rai0018.wav	95,040	51.6212	0.000140	0.0118	2.8554	1.715993
	Rai0020.wav	60,001	30.8188	0.000973	0.0312	2.8003	1.594988
	Rai0001.wav	120,640	70.7971	0.000170	0.0417	2.8875	1.685254
	Rai0002.wav	101,760	60.7208	0.000859	0.0293	2.8652	1.654783

Table 7 Perceptual based analysis

Algorithm	Wav file	No. of Samples	Covl	Csig	Cbak	PESQ	MOS
Proposed 10.6 kbps CS-ACELP	Rai0004.wav	66,400	3.6314	4.2726	2.9967	3.0052	3.3652
	Rai0005.wav	89,920	3.7245	4.3070	3.0301	3.1768	3.5468
	Rai0016.wav	135,360	3.6139	4.2491	2.9648	3.0060	3.3660
	Rai0018.wav	95,040	3.6044	4.1923	2.9954	3.0321	3.3921
	Rai0020.wav	60,001	3.6224	4.2481	2.9697	3.0241	3.3841
	Rai0001.wav	120,640	3.5530	4.1680	2.9328	3.0004	3.3404
	Rai0002.wav	101,760	3.3714	4.0247	2.8118	3.0001	3.3318
Proposed 10.6 kbps CS-ACELP with no searching complexity	Rai0004.wav	66,400	2.5642	2.9922	2.2565	2.3236	2.6736
	Rai0005.wav	89,920	2.3732	2.8034	2.1597	2.2335	2.5835
	Rai0016.wav	135,360	2.6896	3.0687	2.3371	2.4847	2.8247
	Rai0018.wav	95,040	2.5372	2.8876	2.2751	2.4015	2.7415
	Rai0020.wav	60,001	2.3018	2.7095	2.1463	2.1810	2.5410
	Rai0001.wav	120,640	2.5339	2.8861	2.2846	2.3773	2.7173
	Rai0002.wav	101,760	2.5605	2.9080	2.2945	2.4058	2.7358
Legacy 8 kbps CS-ACELP	Rai0004.wav	66,400	2.4529	2.8814	2.2174	2.2170	2.5690
	Rai0005.wav	89,920	2.3828	2.7599	2.2025	2.1625	2.5021
	Rai0016.wav	135,360	2.6200	2.9516	2.3454	2.3980	2.7480
	Rai0018.wav	95,040	2.5112	2.8214	2.2970	2.3256	2.6556
	Rai0020.wav	60,001	2.3037	2.6571	2.1640	2.1001	2.4401
	Rai0001.wav	120,640	3.5530	4.1680	2.9328	2.3725	2.7125
	Rai0002.wav	101,760	3.3714	4.0247	2.8118	2.3411	2.6811

Table 8 Spectral based objective evaluation

Algorithm	Wav file	No. of samples	LLR	WSS	fwSNRseg	ISD	CEP
Proposed 10 kbps CS-ACELP	Rai0004.wav	6640	00.381967	26.612310	8.9778	0.3404	3.8831
	Rai0005.wav	89,920	0.412366	30.806493	8.6113	0.3178	4.0067
	Rai0016.wav	135,360	0.384117	29.030904	8.1853	0.2435	3.8054
	Rai0018.wav	95,040	0.473758	26.83695	7.4741	0.6160	4.5296
	Rai0020.wav	60,001	0.394903	29.122442	8.4652	0.2418	3.9470
	Rai0001.wav	120,640	0.433008	30.063938	8.2623	0.3725	4.1765
	Rai0002.wav	101,760	0.433140	31.978337	7.8659	0.2836	4.1915
Proposed 10.6 kbps CS-ACELP with no searching complexity	Rai0004.wav	66,400	0.928946	60.671137	8.4284	0.7654	4.2072
	Rai0005.wav	89,920	0.988286	62.658947	8.2256	0.6129	4.1099
	Rai0016.wav	135,360	0.958244	60.290989	8.1377	0.9248	4.1341
	Rai0018.wav	95,040	1.061816	60.316124	7.9179	0.4118	4.3399
	Rai0020.wav	60,001	1.060955	60.920406	7.9596	0.6137	4.1794
	Rai0001.wav	120,640	1.059117	60.504075	7.9752	0.5596	4.1306
	Rai0002.wav	101,760	1.047447	61.985622	8.2288	0.7737	4.1121
Legacy 8 kbps CS-ACELP	Rai0004.wav	66,400	0.966210	61.580966	8.4905	0.1149	3.7985
	Rai0005.wav	89,920	1.056541	64.426495	8.3404	0.5065	3.7926
	Rai0016.wav	135,360	1.052315	61.421469	8.3772	0.2254	3.7953
	Rai0018.wav	95,040	1.121723	62.430397	8.0294	0.3188	3.9868
	Rai0020.wav	60,001	1.108372	67.110128	7.9729	0.5479	3.8329
	Rai0001.wav	120,640	0.433008	60.063938	8.2623	0.3725	4.1765
	Rai0002.wav	101,760	0.433140	61.978337	7.8659	0.2836	4.1915

mean of 95% confidence interval. The value of a 1 sample is 3.5468, which is a MOS value of Rai0005.wav,[7] is beyond the range of highest value (3.457) of population mean of 95% confidence interval.

6.4 Searching complexity analysis using simulation delay

In order to compute the searching complexity in terms of simulation delay for a given program in MATLAB, the time required to reconstruct and recover the wave file is calculated using two commands called as 'tic' and 'toc' (Ninad and Kosta 2012). The execution time in the case of proposed 10.6 kbps CS-ACELP based speech codec by considering initial codevector as a final convector is a least with an average of 5.02 s (average of all wave files). While the execution time required for proposed 10.6 kbps is 5.85 s (average of all wave files) and legacy 8 kbps CS-ACELP based speech codec is 5.30 s. The above values are calculated for all mentioned wave files. It can be advocated from the analysis that, with the increase in the bit rate which includes the standard exhaustive search procedure of excitation codevector, the simulation time increases hence it is true in the case of proposed 10.6 kbps speech

Table 9 Calculation of standard deviation for the samples of PESQ

x_i	$\ddot{x}$	$(x_i - \ddot{x})$	$(x_i - \ddot{x})^2$
3.0004	3.0238	−0.0234	0.000548
3.0001	3.0238	−0.0237	0.000562
3.0068	3.0238	−0.017	0.000289
3.0052	3.0238	−0.0186	0.000346
3.1768	3.0238	0.153	0.023409
3.0012	3.0238	−0.0226	0.000511
3.0141	3.0238	−0.0097	0.0000941
3.0008	3.0238	−0.023	0.000529
3.021	3.0238	−0.0028	0.00000784
3.0541	3.0238	0.0303	0.000918
3.0202	3.0238	−0.0036	0.000013
3.0028	3.0238	−0.021	0.000441
3.0001	3.0238	−0.0237	0.000562
3.0654	3.0238	0.0416	0.001731
3.0002	3.0238	−0.0236	0.000557
3.006	3.0238	−0.0178	0.000317
3.0041	3.0238	−0.0197	0.000388
3.0321	3.0238	0.0083	0.0000689
3.0412	3.0238	0.0174	0.000303
3.0241	3.0238	0.0003	0.00000009

codec compared to legacy 8 kbps CS-ACELP based speech codec. The crucial computation is to calculate execution time required when initial codevector is considered

---

[7] See footnote 1.

Table 10 Calculation of standard deviation for the samples of MOS

x_i	$\ddot{x}$	$(x_i - \ddot{x})$	$(x_i - \ddot{x})^2$
3.3404	3.3733	−0.0329	0.001082
3.3318	3.3733	−0.0415	0.001722
3.369	3.3733	−0.0043	0.0000185
3.3652	3.3733	−0.0081	0.0000656
3.5468	3.3733	0.1735	0.030102
3.3412	3.3733	−0.0321	0.00103
3.3741	3.3733	0.0008	0.00000064
3.3508	3.3733	−0.0225	0.000506
3.3887	3.3733	0.0154	0.000237
3.3941	3.3733	0.0208	0.000433
3.3702	3.3733	−0.0031	0.00000961
3.3428	3.3733	−0.0305	0.00093
3.3301	3.3733	−0.0432	0.001866
3.3854	3.3733	0.0121	0.000146
3.3502	3.3733	−0.0231	0.000534
3.366	3.3733	−0.0073	0.0000533
3.358	3.3733	−0.0153	0.000234
3.3921	3.3733	0.0188	0.000353
3.385	3.3733	0.0117	0.000137
3.3841	3.3733	0.0108	0.000117

as a final codevector in the search of excitation codevector which require no exhaustive searches compared to the execution time required in legacy speech coder. From the analysis, it is revealed that the average time required is less when initial codevector is considered as final codevector compared to the time required to final codevector in case of legacy speech coder.

7 Discussion and concluding remarks

CS-ACELP (G.729) is extensively used in VoIP applications which are today's one of the most emerging applications on the smartphones because of its low bit rate requirement to transmit the hybrid traffic through the communication channel.

Basic aim behind implementation of proposed 10.6 kbps CS-ACELP speech coder is to reduce the search engine complexity of excitation codebook structure by introducing excitation codebook structure with a less number of searches compared to legacy coder. Coding complexity is also reduced in comparison with standard CS-ACELP 11.8 kbps speech coder by using the same excitation codebook structure both at the transmitter and receiver which is a different bit excitation codebook structure at transmitter as well as receiver in former. Searching complexity of excitation codebook codevector is also reduced by considering initial codevector as a final codevector.

Results of subjective and objective analysis of proposed CS-ACELP 10.6 kbps speech coder are fairly good compares to legacy CS-ACELP 8 kbps speech coder. The proposed CS-ACELP 10.6 kbps speech coder requires 60 bits per frame for final excitation codevector transmission through the channel, while 11.8 kbps excitation codebook structure requires 70 bits per frame for final excitation codevector transmission through the channel. The proposed coder is a better trade-off option between the two legacy/standard speech coder of CS-ACELP (8 and 11.8 kbps), which provides reduction in number of searches in determining final best optimized excitation codevector compared to the legacy CS-ACELP 8 kbps speech coder, while it also transmits less number of bits for the coding of excitation codevector compared to the requirement of number of bits for coding excitation codevector in standard CS-ACELP 11.8 kbps speech coder.

Efficiency of a proposed algorithm is also evaluated with the population mean of 95% confidence interval (CI95) with the results of objective and subjective quality assessment parameters like PESQ and MOS for different wave files from a standard speech corpus database as an input to proposed CS-ACELP based speech coder. As per the observation, the range calculated for population mean of 95% confidence interval for seven number of samples as individual inputs in terms of PESQ and MOS analysis results, incorporates maximum sample values of subjective or objective quality assessment parameters results which were taken as sample inputs for the calculation of population mean. The observation ensures the consistency of proposed algorithm for different values of PESQ and MOS for different wave files. it is also concluded that the samples which are beyond the scope of range of 95% confidence intervals are the samples having the value greater than the highest value of range of 95% confidence interval, which assure the quite good quality of output decoded speech from the observation of PESQ and MOS ratings (Table 5).

References

Adoul, J. P., Laflamme, C. (1997). Depth-first algebraic codebook search for fast coding of speech, US Patent US5701392A. December 23.

Bernard, A. P. (2005). Algebraic codebook system and Method. US Patent US6847929B2. January 25.

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.

Falk T. H., & Chan, W. (2006). Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 1935–1947

Grundlehner, B., Lecocq, J., Balan, R., & Rosca, J. (2005). Performance assessment method for speech enhancement systems. In *Proceedings of 1st annual IEEE BENELUX/DSP valley signal processing symposium* (pp. 1–4).

Hu, Y., Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Speech Audio Process, 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*.

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

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. US Patent US8185385B2, April 26.

Miao, H., Huang, L., Chen Z., Yang, W., Al-Hawbani, A. (2012). A new scheme for covert communication via 3G encoded speech. *Computers and Electrical Engineering, 38*(6), 1490–1501.

Morgan, D., Permutt, T., & Seldrup, J. (2017). Pharmaceutical statistics. *The Journal of Applied Statistics in the Pharmaceutical Industry*. doi: 10.1002/pst.1783.

Ninad, B., 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.

Salami, R., Laflamme, C., Adoul, J. -P., et al. (1998). Design and description of CS-ACELP: a toll quality 8 kb /s speech coder. *IEEE Transactions on Speech and Audio Processing, 6*(2), 116–130.

Salmela, J., & Mattila, V., (2004). New intrusive method for the objective quality evaluation of acoustic noise suppression in mobile communications. In *Proceedings of 116th audio engineering society convention*. Preprint 6145.

Tahilramani, N., Bhatt, N. (2015). Steganography in speech signal with Enhanced multi-pulse excitation codevector with reduced number of bits. *IEEE international conference on Electrical, Electronics, signals, communication and optimization (EESCO)*, *1–4*, 24–25.

Wu, Z., & Yang, W. (2006). Speech information hiding in g.729. *Chinese Journal of Electronics. 15*(3), 545–549.