Proposed modifications in ITU-T G.729 8 Kbps CS-ACELP speech codec and its overall comparative performance analysis with CELP based 12.2 Kbps AMR-NB speech codec

A. Nikunj Tahilramani, B. Ninad Bhatt

***Abstract*—** This paper proposes the approach of exploiting the use of excitation codebook structure of standard extended G.729 11.8 Kbps [1] having 2 non-zero pulses per track in existing standard 8 Kbps CS-ACELP (80 bits/10 ms) speech codec[1]. Proposed approach avoids the use of two algebraic codebook structure for forward mode as well as for backward mode of G.729E working at 11.8 Kbps with least significant pulse replacement approach for finding optimized excitation codevector. Proposed excitation codebook structure modification in standard 8 Kbps CS-ACELP (80 bits/10 ms) speech codec actuates the bit rate of 11.6 Kbps (116 bits/10 ms) .This paper introduces a comparative analysis between proposed 11.6 Kbps CS-ACELP based speech codec and standard Adaptive multi rate-Narrow band (AMR-NB) 12.2 CELP based speech codec [2]. The comparative analysis shows that results of subjective and objective parameters are quite fair in case of proposed 11.6 Kbps CS-ACELP based speech codec than 12.2 Kbps AMR-NB CELP based speech codec. Here proposed CS-ACELP 11.6 Kbps speech codec is implemented in MATLAB.

Keywords Excitation codebook**,** Excitation Codevector**,** G.729, G.729E, CS-ACELP, ITU-T, MATLAB, Subjective analysis, Objective analysis

I. Introduction

Conjugate structure algebraic-code-excited linear prediction (CS-ACELP) speech coder basically categorized in to Hybrid coder [3] (Analysis by Synthesis coder) classification which provides suitable trade-off between vocoders and waveform coders [3] with satisfactory speech quality and transmission bit rate. Research and development in the domain of source coding techniques like CELP (Code Excited Linear Prediction), ACELP (Algebraic CELP) and CS-ACELP (conjugate structure-Algebraic CELP) still continue to emerge as a famous area of research worldwide [3].

The authors are with the Silver Oak College of Engineering & Technology, Ahmedabad, INDIA

Input to the speech encoder is 16-bit PCM from the audio part of the mobile station terminal. Before encoding, the 64 Kbit/s data, should be converted to 16 bit linear PCM and 16 bit linear PCM to the suitable form after synthesis. 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 [1]. Adaptive multi rate codecs are used by GSM and UMTS [2]. It is totally handled by radio access network. During heavy call traffic it switches to lower bit rates to provide higher capacity but slightly lower speech quality. The tradeoffs between the speech quality, bit rates and coverage can be achieved with AMR codecs. AMR works on 20 ms speech frames with 160 samples per frame with 8000 samples/second sampling frequency.

This paper is organized as follows. In Sect. 2, CELP based GSM AMR-NB coder is introduced. In Sect. 3, excitation sequence of CS-ACELP based coder is explained. In Sect. 4 modified fixed codebook structure and the final codevector searching procedure is introduced. In Sect 5, various subjective and objective quality assessment parameters are defined. In Sect 6, comparative performance evaluation of proposed CS-ACELP speech coder and CELP based AMR-NB speech codec is demonstrated using set of graphs and tables. Finally, the concluding remarks are given in Sect. 7.

II. CELP based GSM AMR-NB speech codec

It is standardized CELP based speech codec by 3GPP [2] which switches the bit rate according to channel conditions and background noise. Adaptive multi rate is a standard compression for audio which is utilized for speech coding. AMR is used to encode the narrowband signal with 8 different variable bit rates which ranges from 4.75 Kbps to 12.2 Kbps. The best codec mode out of is selected based on channel to interference ratio(C/I) ratio. Depending on the channel capacity and radio conditions the trade-off decision between source coding and channel coding is being made adaptively by the AMR system. GSM AMR-NB working at 12.2 Kbps is compatible with ETSI GSM enhanced full rate speech codec which was developed to improve the speech quality of GSM-Full rate codec. AMR facilitates voice activity detection (VAD) to detect the voice activity and to generate the signal of discontinuous transmission (DTX) if voice activity is not present to reduce the bandwidth of a channel and hence increases the battery lifetime. It also provides the confirmation regarding link activity during the silence period using comfort noise generator (CNG). AMR-NB speech codec is basically having 8 different bit rates of 12.2, 10.2, 7.95, 7.4, 6.7, 5.9, 5.15 and 4.75 Kbps.

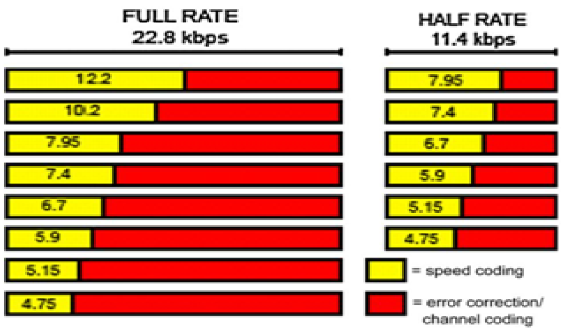

Figure: 1 AMR full-rate and half rate codec [2]

III. CS-ACELP Speech codec excitation sequence generation

3.1 Modification in standard excitation codebook structure of 8 Kbps CS-ACELP based speech codec

Standard CS-ACELP based speech codec working at 8 Kbps is having 40 pulse positions with first three tracks are having 8 positions each and the final track is incorporating 16 pulse positions. The final excitation codevector is determined by different optimizations and recursive number of searches by considering different pulse contribution from different tracks at the starting of exhaustive searching procedure. In [7] , search engines requires **8192** no. of searches in case of full search approach, search engine needs **1440** no. of searches in case of focused search approach and [7] determines the final excitation codevector with **320** no. of searches in case of Depth first search approach.

As the final track of excitation codebook structure consist of 16 positions in standard 8 Kbps CS-ACELP based speech codec , it requires to consider first 8 positions of final track while considering 8 positions from each track in search engine while determining best excitation codevector. In the proposed approach the excitation structure of standard 8 Kbps CS-ACELP based speech codec is replaced with (Table 3) the extended 11.8 Kbps CS-ACELP based speech codec forward mode excitation codebook structure (having 5 tracks with each track is having 8 pulse positions.

Table 1 Bit allocation of the 11.6 kbit/s CS-ACELP algorithm (10 msec frame)

Parameter	**Subframe1**	**Subframe 2**	**Total bits/ frame**
LSP	-	-	18
Adaptive-codebook delay	8	5	13
pitch-delay parity	1	-	1
Fixed- codebook index	30	30	60
Fixed- codebook Sign	5	5	10
Codebook gain (stage 1)	3	3	6
Codebook gain (stage 2)	4	4	8
Total			116

Final codevector is made up of 10 pulse positions with contribution of 2 best pulse positions from each track. As there are two pulse positions which are selected per track it require 6 bits per track for transmission of the codebook index and 1 sign bit is required per track to indicate the sign of the selected 2 pulse positions having sign magnitude ±1. Total 35 bits are transmitted per subframe for excitation codevector position and its sign.

Table 2 Fixed codebook excitation structure of standard 11.8 Kbps CS-ACELP based Speech Codec in forward LP mode[1]

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

IV. Searching procedure of excitation codevector in proposed approach

First pulse position form each track among the two from each individual track is found by the maximum of the correlation vector d(n) of the 8 positions of a particular track and second pulse in each track is determined by finding the second maximum with respect to first maximum of each individual track. Least significant pulse replacement approach is used to determine the best combination of codevector from all five tracks to determine final excitation codevector. After performing pulse replacement the pulses which maximizes the value of Qk (eq.1) among all 60 Qk searches with 12 Qk searches per track will be kept for the final excitation codevector which in turn minimizes the residual error [5].

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

Here M is a number of tracks in a subframe analysis.

A K[th] 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[5]:

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

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

From Eq.2 and Eq.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). Also a numerator and a denominator of Eq.1 are described as [5]:

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

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

Number of pulses in sub-frame is described as Np and m denotes a position of i[th] pulse. Along with the 2 pulse positions which are coded with 3 bits individually as total 8 positions are there in each track, the sign of only pulse position is transmitted in terms of sign magnitude of ±1. The sign of the second pulse at the receiver is determined from the sign of the first pulse itself from the presumption of the position of the pulse selected. The smallest position among the two is identified as p1 and other as p2 with their respective signs s1 and s2. At the receiver if the first decoded pulse position is less than or equal to second then sign of both pulses are same and which is denoted by the sign bit which is transmitted as +1 or -1 otherwise the sign of the two pulses are different. If the sign of the two pulses are different then the sign of the first pulse is transmitted which is a largest among the two pulse positions and second pulse is having the opposite sign then that of first once it is decoded at the receiver.

4. Subjective and objective measures

Overall performance of the proposed codec is evaluated by subjective and objective quality assessment parameters. Subjective measure is categorized into MOS (Mean opinion score) and objective measure is categorized into waveform based, spectral based and perceptual based analysis.

a. Subjective measures

The quality of the compressed speech is determined by MOS assessment. In subjective measure, quality of the output speech is asked to judge by 5 to 10 subjects and they are asked to rate the output compressed speech quality according to the choices given in a Table 5.

b. Objective measures

Objective measures are further classified into waveform based, spectral based, perceptual based and composite measure based analysis [],

i. Waveform based analysis

Following quality assessment parameters are evaluated in above classification [2],

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

$$AbsErr = \sum |S_i - S_o| \quad (6)$$

(2) Mean Square Error (MSE) is mathematically expressed as

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

(3) Root Mean Square Error (RMSE) is mathematically expressed as

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

(4) Signal to Noise Ratio is mathematically given as,

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

Where *Si*=input signal, So=decoded output signal and N=total no. of frames.

(5) Segmental SNR is mathematically given as

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

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.

ii. Perceptual based analysis

Perceptual evaluation of speech quality (PESQ):
PESQ algorithm uses psychoacoustic and cognitive models. By using a synchronization scheme, this algorithm time aligns the original and degraded speech signals, as misalignment could result in a false quality score. PESQ is designed to analyze specific parameters of audio, including time warping, variable delays, transcoding, and noise. PESQ score is calculated as a liner combination of the average disturbance value Dind and the average asymmetrical disturbance value Aind as [2]:

$$PESQ=a_0+a_1Dind+a_2Aind \quad (10)$$

Where a0, a1 and a2 are calculated using Multiple linear regression analysis.

iii. Spectral based analysis

Following parameters are categorized to perform spectral based analysis[2].

1. Log Likelihood Ratio (LLR) is calculated by following mathematical formula:

$$d_{LLR}(\vec{a_p},\vec{a_c}) = \log\left(\frac{\vec{a_p}\,R_c\,\vec{a_p^T}}{\vec{a_c}\,R_c\,\vec{a_c^T}}\right) \quad (11)$$

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

2. Itukara Saito Distance measure is mathematically defined as

$$d_{IS}(\vec{a_p},\vec{a_c}) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a_p}\,R_c\,\vec{a_p^T}}{\vec{a_c}\,R_c\,\vec{a_c^T}}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \quad (12)$$

Where σ_p^2 and σ_c^2 are LPC gains of original and decoded signals [2]. 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:

$$c(m) = a_m + \sum_{k=1}^{m-1}\frac{k}{m}c(k)a_{m-k} \quad (13)$$

Where p denotes the order of the LPC analysis. An objective measurement from the Cepstrum coefficient can be computed with following expression:

$$d_{CEP}(\vec{c_c},\vec{c_p}) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}[c_c(k)-c_p(k)]^2} \quad (14)$$

Where $\vec{c_c}$ and $\vec{c_p}$ 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) $fwSNR_{seg}$

$$= \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)\log_{10}\frac{|X(j,m)|^2}{\left(|X(j,m)|-|\hat{X}(j,m)|\right)^2}}{\sum_{j=1}^{K}W(j,m)} \quad (15)$$

Where W(j,m) is denoted as weight placed on the j^th 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 j^th frequency band at the m^th frame, while $|\hat{X}(j,m)|$ is denoted as weighted decoded signal spectrum in the same band.

5. Weighted slop spectrum distance is defined as $fwSNR_s$

$$(\overline{11}\;\frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)\log_{10}(|s_c(j,m)|-|s_p(j,m)|)^2}{\sum_{j=1}^{K}W(j,m)} \quad (16)$$

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 j^th frequency band at frame m of the original and decoded speech signal with total of 25 no. of bands[2].

iv. Composite measures

Unlike the simple objective measures parameters, there are certain parameter which combines 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 [2][3][6]:

Csig= 3.903 - 1.029. LLR+ 0.603. PESQ - 0.009. WSS (17)

Cbak= 1.634 + 0.478 . PESQ – 0.007 . WSS + 0.063. segSNR (18)

Covl=1.594 + 0.805 . PESQ – 0.512 . LLR - 0.007 . WSS (19)

V. Simulation of proposed algorithm and comparative analysis with CELP based AMR-NB 12.2 Kbps speech codec based on MATLAB

Here proposed CS-ACELP based 11.6 Kbps speech codec is implemented in Matlab and its performance results in terms of different objective and subjective parameters are compared with 12.2 Kbps AMR-NB CELP based speech codec. For the comparative analysis four different wav files have been chosen NOIZEUS (The NOIZEUS database 2009) corpus. Out of four different wav files two are of female and other two are of male speakers. All the wav files are having a sampling frequency of 8000 samples/second and each sample is represented by 16 bits/sample which results into 128 Kbps bitrate.

A. Result obtained for MOS analysis

For subjective analysis 10 different un-trained subjects are identified to take part in MOS ratings. Out of 10 , 5 subjects are men and 5 are women. Each subject is offered with 8 different wav files. The judgement given by all 10 subjects in form of rating is averaged to calculate the final MOS ratings. The comparative results of the MOS score of proposed 11.6 Kbps CS-ACELP based speech codec and 12.2 Kbps.

AMR-NB CELP based speech codec are shown in fig 2. As it can be observed from the figure that MOS score of proposed CS-ACELP based speech codec is far better than the CELP based speech codec. The quality, intelligibility and pleasantness of the proposed codec decoded speech is quite good compare to GSM AMR based speech codec [] though the bit rate is less of proposed speech codec.

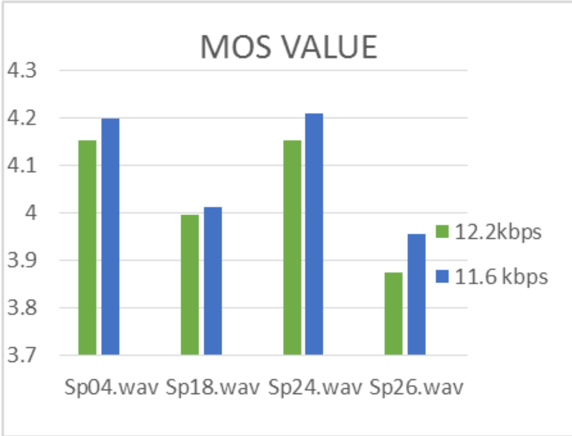

Figure: 2 MOS score for different wave files

B. Results obtained for objective analysis

The results obtained for objective analysis for three different categories are shown in the table 3, 4 and 5.

1. Waveform based analysis

It can be observed from the Table 3 that results of the absolute error, mean square error, root mean square error, signal to noise ratio and segmental signal to noise ratio in case proposed CS-ACELP based speech codec are quite comparable with existing CELP based AMR-NB 12.2 Kbps are quite fair, though proposed coder is having a short fall of 6 bits as it is reducing the bandwidth compare to existing coder. Moreover decrease in the bit rate form 12.2 Kbps to 11.6 Kbps increases the results of the quality assessment parameters.

Table 3 Waveform based objective evaluation

GSM AMR 12.2 kbps					
Wave file	ABS	MSE	RMSE	SNR	SNRseg
Sp04.wav	99.203	0.000113	0.0106	12.6828	8.4478
Sp18.wav	101.63	0.000112	0.0106	11.5163	7.6934
Sp24.wav	144.67	0.000083	0.0091	12.6091	10.2304
Sp26.wav	136.53	0.000146	0.0121	10.8590	8.2569
Proposed CS-ACELP (11.6 kbps)					
Wave file	ABS	MSE	RMSE	SNR	SNRseg
Sp04.wav	0.6131	0.000100	0.0104	12.6685	8.5519
Sp18.wav	1.5637	0.000102	0.0105	11.7989	7.8541
Sp24.wav	0.4278	0.000079	0.0087	0.0087	10.4882
Sp26.wav	0.3827	0.000138	0.0118	10.9513	8.2591

2. Composite and perceptual based analysis

Figure 4 shows the comparative perceptual based analysis for four different wave files. It can be observed that the results of PESQ for proposed coder are far better than the existing AMR-NB speech codec.

As an individual performance analysis of proposed coder, it can be observed from table 6 that as PESQ decreases the results of Covl, Cbak and Csig also decreases. As a comparative analysis the results of the proposed coder are quite satisfactory.

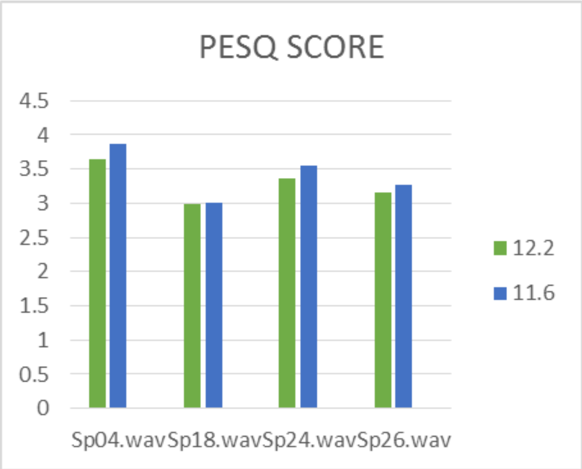

Figure: 3 PESQ score for different wave files

Table 4 Perceptual based analysis and composite measures

GSM AMR 12.2 kbps					
Wave file	Covl	Csig	Cbak	PESQ	MOS
Sp04.wav	4.19	4.75	3.71	3.6500	4.1526
Sp18.wav	3.67	4.39	3.32	2.9971	3.9959
Sp24.wav	3.85	4.35	3.68	3.3640	4.1529
Sp26.wav	3.84	4.53	3.48	3.1568	3.8754
Proposed CS-ACELP (11.6 kbps)					
Wave file	Covl	Csig	Cbak	PESQ	MOS
Sp04.wav	4.22	4.76	3.74	3.8600	4.2000
Sp18.wav	3.70	4.44	3.42	3.0121	4.0121
Sp24.wav	3.88	4.38	3.54	3.5487	4.2100
Sp26.wav	3.86	4.56	3.50	3.2658	3.9547

3. Spectral based analysis

The spectral based parameters like CEP, ISD, WSS, LLR, and fwSNRseg gives the fair values for proposed codec. The results of the proposed codec are quite comparable with the existing GSM-AMR codec.

Table 5 Spectral based objective evaluation

GSM AMR 12.2 kbps					
Wave file	LLR	WSS	fwSNRseg	ISD	CEP
Sp04.wav	0.2783	27.506	27.506	0.3563	2.9008
Sp18.wav	0.2032	32.681	11.7186	0.2968	2.4508
Sp24.wav	0.4941	28.431	14.0519	0.5601	4.8662
Sp26.wav	0.2258	25.203	13.9092	0.2954	2.7915
Proposed CS-ACELP (11.6 kbps)					
Wave file	LLR	WSS	fwSNRseg	ISD	CEP
Sp04.wav	0.3689	27.617	14.8489	0.4080	3.8094
Sp18.wav	0.3108	36.276	10.7227	0.2975	3.4630
Sp24.wav	0.6328	29.837	14.2603	0.9431	5.7174
Sp26.wav	0.3239	26.235	13.9509	2.2962	3.6816

VI. Concluding remarks

As it is a world of VoIP, the recent technologies are extensively doing the research on VoIP based applications. CS-ACELP based speech codecs are most popular speech codecs for VoIP based applications where it is required to transmit voice traffic as well as data traffic. Today the main constraint in communication world is the bandwidth. Idea behind the implementation of proposed CS-ACELP based speech codec working at 11.6 Kbps is to introduce a codec which avoids the use of two codebooks excitation structure used in extended G.729E CS-ACELP based speech codec for determination of optimized excitation codevector as well as it can produce the better decoded speech quality compare to the existing CELP based AMR-NB speech coded working 12.2 Kbps speech codec. Tradeoff between CELP and CS-ACELP is shown with reduction of the bit rate which actuates a new speech codec with less bit rate and good speech quality.

REFERENCE

[1] ITU-T. Recommendation 2007, “coding of speech at 8 kbit/s using conjugate-structure algebraic-cod-excited linear prediction (CS-ACELP)”.

[2] Bhatt N, Kosta Y. Overall performance evaluation of adaptive multi rate 06.90 speech codec based on code excited linear prediction algorithm using MATLAB. Int J Speech Technol, 2012, 15: 119–129

[3] N. Bhatt and Y. Kosta, “Proposed modifications in ETSI GSM 06.10 Full Rate speech codec and its overall evaluation of performance using MATLAB” ”, Int. J. of Speech Tech., Springer, 14, (3, 157, Sept. 2011).

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

[5] Alexis P. Bernard, “Algebraic codebook system and Method”. U.S. Patent US6847929B2, January 25, 2005.

[6] Bhatt Ninad,Kosta Y., “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, 2012

[7] N. Tahilramani, N. Bhatt., “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), pp. 1-4,24-25 Jan 2015