Quality Enhancement of Low Bit Rate Speech Coder with Nonlinear Prediction

Ancy S. Anselam, Sakuntala S. Pillai, and K. G. Sreeni

Abstract Toll quality speech codec design with a low bit rate is really a challenging task in modern communication because of the drastic increase in end-users in social networks. Most of the low bit rate speech codecs are based on linear prediction. The code-excited linear prediction codec (CELP) gives good quality decoded speech at a lower bit rate of 4.8 Kbps. But, it neglects the natural nonlinear effects present in speech production process. So, some adaptive techniques are to be used to make the system nonlinear to perform better than linear prediction speech codecs. An adaptive technique with nonlinear prediction of speech, based on truncated Volterra series, is used to generate the nonlinear prediction coefficients. The generated nonlinear prediction coefficients are implemented in G723.1 CELP codec to introduce code-excited nonlinear prediction (CENLP) codec. Advancements in the performance are evaluated using subjective and objective quality measures and compared with the normal G723.1 CELP codec.

Keywords CELP · Nonlinear long-term prediction · Pitch period · Prediction gain · Volterra series

1 Introduction

To accelerate the efficient transmission and compact storage capabilities, some systems need low bitrate and low delay speech coders with very good quality. Commonly

A. S. Anselam (✉) · S. S. Pillai
Department of Electronics and Communication Engineering, Mar Baselios College of Engineering and Technology, Thiruvananthapuram, India
e-mail: ancy.anselam@mbcet.ac.in

S. S. Pillai
e-mail: sakuntala.pillai@gmail.com

K. G. Sreeni
Department of Electronics and Communication Engineering, College of Engineering Trivandrum, Thiruvananthapuram, India
e-mail: sreenikg79@gmail.com

J. Jayakumari et al. (eds.), *Advances in Communication Systems and Networks*, Lecture Notes in Electrical Engineering 656,
https://doi.org/10.1007/978-981-15-3992-3_53

Table 1 Values of quality measures and delay of existing LPC-based speech codecs

Linear prediction codecs parameters	MELP 2.4 Kbps	CS-ACELP 8 Kbps	LD_CELP 16 Kbps	CELP 4.8 Kbps
Delay (ms)	122.5	10	0.625	37.5
MOS score	2.9	3.98	3.95	3.15

used speech coding algorithms are based on linear prediction. In this, the current sample is predicted using future or past speech samples. Various speech coding algorithms with linear prediction are linear prediction coder [1], code-excited linear prediction (CELP) [2], low delay CELP [3], mixed excitation linear prediction (MELP) [4] and conjugate structure—Algebraic CELP [5]. Among these speech coders, the CELP 4.8 gives the moderate quality and delay. The trade-off between the quality measure (MOS score) and delay of LPC-based speech codecs is shown in Table 1. It shows that to achieve toll-quality decoded speech with low bitrate and low delay alterations in CELP 4.8 are essential.

During the development of speech production models, it is assumed that, a power source (i.e. lungs) produces a puff of air (i.e. sound) and the surroundings (i.e. vocal tract) shapes or filters the same before it reaches the lip end [6]. Linear prediction model works with an assumption that the lungs and the vocal tract are independent, which allows the separation of the source and filter in the model (i.e. source-filter model). It removes the interdependence of the excitation source and the vocal tract. But, this assumption will not work well for all speech processing applications. Therefore, adaptive procedures that consider the nonlinear generation process (dependency lungs and vocal tract) of the speech signal need to be considered, which may outperform the linear techniques, enabling better performance speech processing applications.

A variety of procedures dealing with nonlinear speech prediction are reported in the literature. Among them, the most widely used technique is nonlinear adaptive filtering using Volterra series model [7]. This deals with the short-term and long-term prediction (LTP) of speech based on truncated Volterra series. In this work, the nonlinear characteristics of speech signal are modelled using the Volterra series for short-term and long-term signals [6]. Then, the nonlinear predictor coefficients are obtained using the adaptive techniques. The adaptive technique used is 'recursive least squares (RLS) algorithm' which offers fast convergence and good error performance.

2 Code-Excited Linear Prediction Speech Codec (CELP)

CELP is viewed as the upgraded version of LPC model, in which codebooks used are vector quantization based. It follows analysis-by-synthesis approach [2]. The formant synthesis filter function is:

$$H_f(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}} \tag{1}$$

$A(z)$—the system function of analysis filter and M—order of the filter. The amount of noise in the reconstructed speech is adjusted by using a weighing filter. The system function used for the implementation of perceptual weighing filter is:

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 + \sum_{i=1}^{M} a_i z^{-i}}{1 + \sum_{i=1}^{M} a_i \gamma^i z^{-i}} \tag{2}$$

where γ—constant with values [0, 1]. If it is sampled at 8000 samples per second, γ is between 0.8 and 0.9. Best matching vector is selected from codebook by using MMSE criterion [8].

Figure 1 shows CELP encoder design flow. The speech in PCM format is given to the encoder [2]. Linear prediction is done in two different levels—short-term and long-term. To obtain formant information, short-term prediction is done on speech frames and long-term prediction is done on sub-frames to get pitch and intensity. Finally, CELP bitstream formed by encoding the linear prediction coefficients (LPCs), long-term LP parameters, gain and codebook excitation index.

Figure 2 shows the design flow of CELP decoder. From the received bitstream, parameters are decoded and extracted. The decoder takes the code vector from the codebook corresponding to the index received. The obtained code vector is scaled by the matching gain. The decoder filters the scaled code vector by pitch synthesis and formant synthesis filters with the help of received parameters. Then, the synthesized speech is passed through the post-filter to improve the perceptual quality.

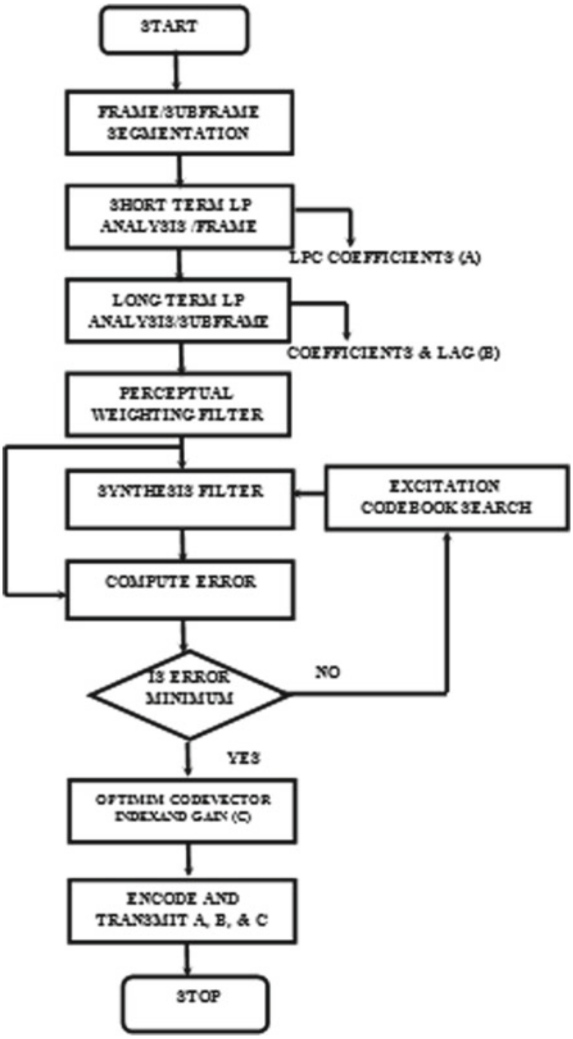
3 Estimation of Nonlinear Prediction Coefficients Using Volterra Series

For the approximation of nonlinear behaviour in systems, the Volterra series model is used similar to Taylor series. It has the ability to capture 'memory' effects, and this makes the model different from Taylor series. The input–output relation of discrete-time Volterra series with infinite memory is as follows [6, 7].

$$\begin{aligned} y(n) &= h_0 + \sum_{k=0}^{\infty} h_1 \,.\, s(n-k) + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_2(i, j) \,.\, s(n-i) \,.\, s(n-j) \\ &\quad + \sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} h_2(i, j, k) \,.\, s(n-i) \,.\, s(n-j) \,.\, s(n-k) + \cdots \end{aligned} \tag{3}$$

where $s(n)$—input signal, $y(n)$—output of the model, h_0—bias coefficient, h_1—linear coefficients, h_2—quadratic coefficients, h_3—cubic coefficients.

Fig. 1 CELP encoder


START
FRAME/SUBFRAME SEGMENTATION
SHORT TERM LP ANALYSIS /FRAME
LPC COEFFICIENTS (A)
LONG TERM LP ANALYSIS/SUBFRAME
COEFFICIENTS & LAG (B)
PERCEPTUAL WEIGHTING FILTER
SYNTHESIS FILTER
EXCITATION CODEBOOK SEARCH
COMPUTE ERROR
IS ERROR MINIMUM
NO
YES
OPTIMUM CODEVECTOR INDEXAND GAIN (C)
ENCODE AND TRANSMIT A, B, & C
STOP

Here, the output is expressed as the linear combination of nonlinear functions of the input signal. So, the adaptation algorithms valid in linear case can be extended to Volterra series. Both LMS and RLS are used to identify unknown parameters. With the order of nonlinearity, the number coefficients of increases exponentially resulting in a significant increase in system complexity. So, the Volterra series is truncated to low linearity orders.

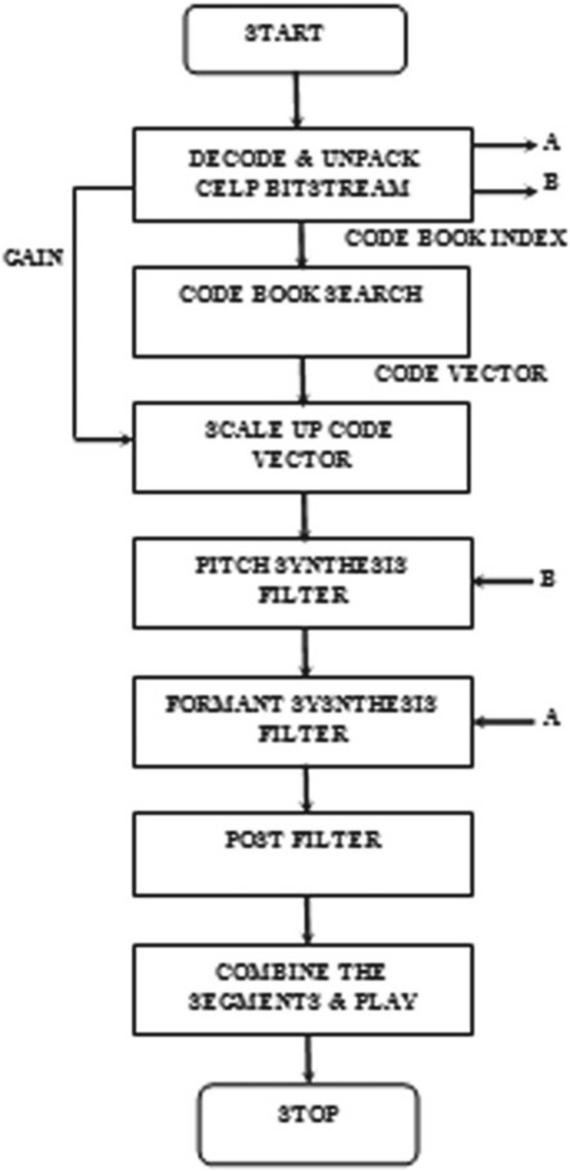


Fig. 2 CELP decoder

3.1 Short-Term Speech Prediction: Second-Order Volterra Model

Volterra series model-based nonlinear predictor estimates the current signal value as a linear combination of past signal values and linear combinations of products of past signal values. The second-order Volterra model treats the system with the first- and second-order kernels only and the predicted signal is given by

$$\hat{s}(n) = h_0 + \sum_{k=1}^{P} h_1(k)s(n-k) + \sum_{i=1}^{M}\sum_{j=1}^{M} h_2(i, j)s(n-i)s(n-j) \tag{4}$$

where $\hat{s}(n)$ is the estimate of $s(n)$. This model consists of two parts: a linear part of prediction order P with coefficients $h_1(k)$ and a quadratic part of prediction order M with coefficients $h_2(i, j)$. With the coefficient symmetry, i.e. $h_2(i, j) = h_2(j, k)$, the overall number of coefficients for the second-order Volterra predictor is

$$n_c = P + \frac{M(M+1)}{2} \tag{5}$$

The system model of a second-order Volterra-based nonlinear predictor with prediction order $P = 2$ is given in Fig. 3.

Prediction error for the filter model is,

$$e(n) = s(n) - \hat{s}(n) = s(n) - h_0 - \sum_{k=0}^{P} h_1(k)s(n-k)$$
$$-\sum_{i=0}^{M}\sum_{j=0}^{M} h_2(i, j)s(n-i)s(n-j) \tag{6}$$

The filter coefficients are computed by minimizing the criterion function based on RLS algorithm,

$$J(n) = \sum_{k=0}^{n} \rho^{n-k} \,.\, \mathrm{e}^2(k) \tag{7}$$

where $e(n)$—prediction error and ρ—forgetting factor ($0 \ll \rho \leq 1$).

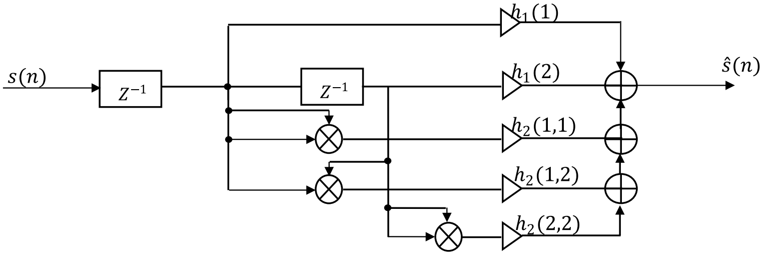


Fig. 3 Second-order Volterra filter model

3.2 Long-Term Speech Prediction Using Volterra Model

Only the correlation between nearby samples is eliminated by short-term linear prediction. To model voice signal adequately, the prediction order must be high enough to include at least one pitch period. Due to large delay and increased complexity, this is not acceptable in most of the practical applications. Hence, it is really important to design long-term predictor (LTP) that is capable of removing far-sample redundancies due to the presence of a pitch excitation. The standard solution to this problem in linear prediction is the use of a model with two linear predictors, short-term and long-term, connected in cascade as shown in Fig. 4. This is used in many of the linear prediction speech coders [2].

Second-Order Long-Term Volterra Prediction To perform the nonlinear long-term prediction of speech signal, a short-term linear predictor and a long-term second-order Volterra predictor are connected in cascade as shown in Fig. 5. This filter model predicts the current speech sample from a past sample which is one or more pitch periods away.

The predicted sample from the above filter model is

$$\hat{e}_s(n) = h_1 \, . \, e_s(n - T) + h_2 \, . \, e_s^2(n - T) \tag{8}$$

where T—pitch period and h_1 and h_2—long-term prediction coefficients. For a given T, the coefficients are computed by minimizing the sum of squared error.

Sum squared error is given as,

$$J = \sum_n \big(e_s(n) - \hat{e}_s(n)\big)^2 \tag{9}$$

The long-term prediction coefficients obtained are,

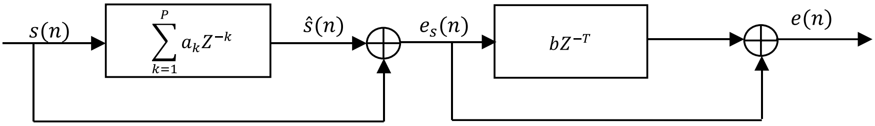


Fig. 4 Short-term and long-term linear predictors in cascade

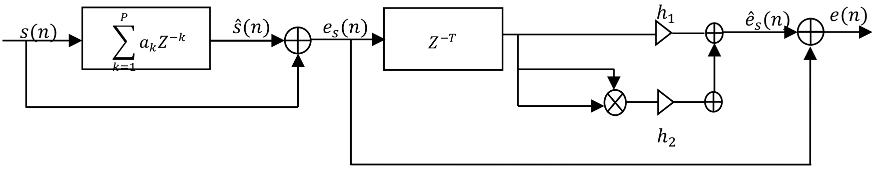


Fig. 5 Short-term predictor and second-order long-term Volterra predictors in cascade

$$h_1 = \frac{\sum_n e_s^4(n-T)\sum_n e_s(n)\,.\,e_s(n-T) - \sum_n e_s^3(n-T)\sum_n e_s(n)\,.\,e_s^2(n-T)}{\sum_n e_s^4(n-T)\sum_n e_s^2(n-T) - \left(\sum_n e_s^3(n-T)\right)^2} \tag{10}$$

$$h_2 = \frac{\sum_n e_s^2(n-T)\,.\,\sum_n e_s(n)\,.\,e_s^2(n-T) - \sum_n e_s^3(n-T)\,.\,\sum_n e_s(n)\,.\,e_s(n-T)}{\sum_n e_s^4(n-T)\sum_n e_s^2(n-T) - \left(\sum_n e_s^3(n-T)\right)^2} \tag{11}$$

4 Code-Excited Nonlinear Prediction Speech Coder (CENLP)

The CENLP–4.8 coder uses the analysis-by-synthesis method of encoding the signal, i.e. the encoder synthesizes the signal (without any channel errors). The CENLP encoder and decoder flow diagrams are shown in Fig. 6.

Short-term linear prediction: Estimation of the formant structure (the period of the formants is less than 1 ms) and the nonlinear prediction coefficients of different orders like 2, 4, 10. These nonlinear coefficients are used to find the error signal by the adaptive technique, which enhance the quality of speech signal.

Long-term linear prediction: Estimation of the pitch and intensity of the speech signal (the period of these parameters is in the order of 3–15 ms). The degree of nonlinearity is maximum up to 3, since the complexity in structure makes it computationally difficult to find the coefficients. Here the long-term predictions up to orders 2 and 3 are done to found the coefficients, and thus, the error signal is determined and the signal is reconstructed for the quality checking purpose.

The nonlinear prediction coefficients (NLPCs) are given as input to the prediction-error filter which performs inverse filtering, i.e. the formant information provided by the NLPCs is removed from the speech sub-frame signals, giving the resultant signal from which long-term parameters like pitch and intensity can be estimated (by estimating the frequency and amplitude of the signal). The pitch and intensity are the long-term parameters that are sent for encoding to be transmitted. The input frame is also given as input to a perceptually weighted filter. The filter weights are determined based on a perceptual compression algorithm. The short-term prediction output (i.e. NLPCs) and the long-term prediction output (pitch and intensity) are given as input to formant synthesis filter and pitch synthesis filter, respectively.

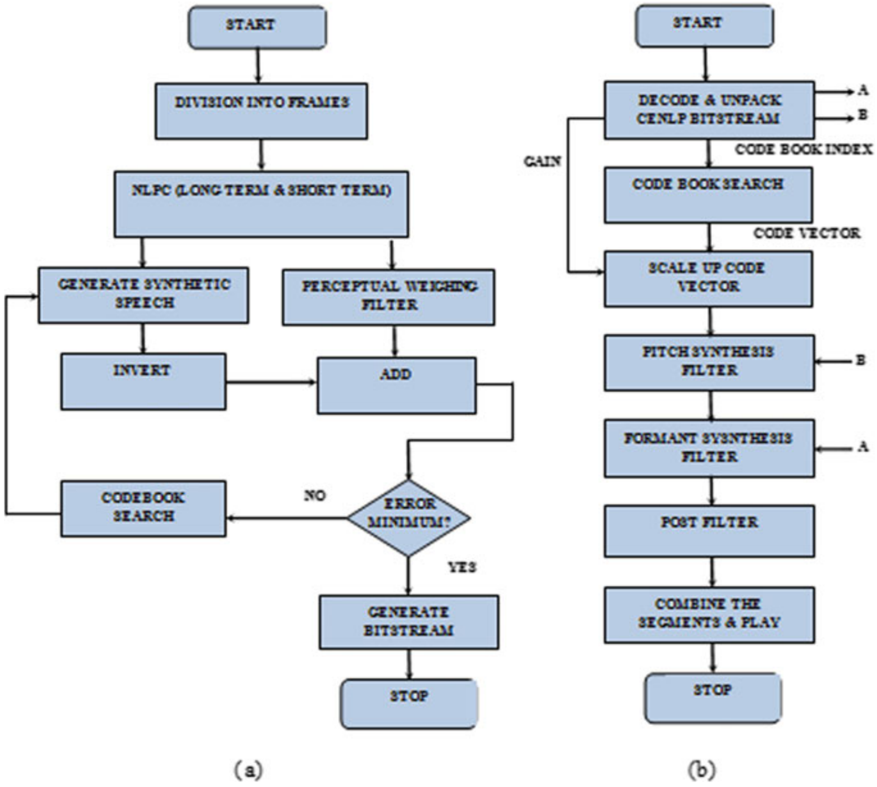


Fig. 6 **a** Design flow of CENLP Encoder **b** Design flow of CENLP Decoder

4.1 Implementation Details of CENLP Coder with Bitrate 4.8 Kbps

CENLP codec is implemented in MATLAB for a bit rate 4.8 Kbps. Speech sampled at 8 kHz is given to the encoder. The speech is divided into frames of 240 samples each (30 ms). The number of frames decides the number of times the encoder must run. Then it is subdivided into sub-frames of 60 samples each (75 ms).

The frames are given to the second-order Volterra short-term nonlinear predictor to compute the nonlinear prediction coefficients. The filer used to have the linear prediction order $P = 4$ and nonlinear prediction order $M = 3$. The output of this block is an array of coefficients 1, $h_1(1)$, $h_1(2)$, $h_1(3)$, $h_1(4)$, $h_2(1,1)$, $h_2(1,2)$, $h_2(1,3)$, $h_2(2,2)$, $h_2(2,3)$ and $h_2(3,3)$. These coefficients are nonlinear prediction coefficients (NLPCs).

The predicted speech sample is given by the equation,

Table 2 Variation in prediction gain

Type	Prediction gain in dB
Short-term LPC ($P = 10$)	12.5
Short-term LPC and long-term linear prediction	14.8
Short-term LPC and Volterra long-term prediction	17.6

$$\hat{s}(n) = 1 + \sum_{k=1}^{4} h_1(k)s(n-k) + \sum_{i=1}^{3}\sum_{j=1}^{3} h_2(i, j)s(n-i)s(n-j) \qquad (12)$$

To find the periodicity of voiced speech frames, a long-term predictor combined with a short-term predictor is used. The long-term prediction is done at the sub-frame level. For this, a cascade of linear short-term predictor and a second-order Volterra long-term nonlinear predictor is used. Short-term prediction error appears as the input long-term predictor. Through the squared error minimization procedure, long-term prediction coefficients (LTPCs) are computed. Then, a search procedure is done within a range $T_{\min} \leq T \leq T_{\max}$ to determine the pitch period (T) [6]. Typically, $T_{\min} = 20$ ms and $T_{\max} = 140$ ms.

Sub-frame level nonlinear predictions are done under different conditions, and variation in prediction gain is listed in Table 2.

Parameters considered for the encoding of CENLP 4.8 codec are as follows.

Number of samples/s	8000
Frame duration	60 ms
Frame length	240 samples
Number of nonlinear prediction coefficients (NLPCs)	10/frame ($P = 4$ and $M = 3$)
Pitch filter coefficients (LTPCs)	2/sub-frame
Number of gain values	4/frame
Pitch filter delays	4/frame
Codebook size	512 code vectors (index length–9).

These parameters are encoded as per the following bit allocation strategy.

NLPC coefficients—6 × 10 = 60, gain—2 × 4 = 8, Pitch filter coefficients—3 × 8 = 24, delay—4 × 4 = 16 and codebook index = 4 × 9 = 36.

5 Experimental Results

Both the nonlinear and linear predictions were implemented in the G723.1 CELP codec, and the results were observed. The reconstructed signals had very good quality

Table 3 MOS score CELP and CENLP

Language	File type	MOS score	
		CELP	CENLP
Malayalam	Mal_female.wav	3.68	3.56
	Mal_male.wav	3.45	3.85
	Mal_Conv.wav	2.85	3.55
English	Eng_female.wav	4.34	4.12
	Eng_male.wav	4.78	4.55
	Eng_Conv.wav	3.80	3.98
Hindi	Hin_female.wav	4.12	4.35
	Hin_male.wav	3.45	4.26
	Hin_Conv.wav	3.64	3.75
Tamil	Tam_female.wav	3.45	3.75
	Tam_male.wav	4.10	4.00
	Tam_Conv.wav	3.89	3.95
Kids	Hin_female.wav	3.40	3.89
	Hin_male.wav	2.98	3.52
Average		3.694	3.92

compared to the original speech signal. The subjective testing and objective testing were performed on the reconstructed speech signals.

5.1 Subjective Testing

Subjective evaluation of in CELP and CENLP codecs was done by conducting informal listening test (MOS testing) of reconstructed speech [11]. It was conducted using the speech.wav files of the database created for Indian dialects. The results of the MOS testing of coder outputs for several speech signals are included in Table 3. The MOS score obtained from reconstructed speech files for CENLP codec is 3.92, which approaches toll quality.

5.2 Objective Testing

Objective distortion measures of codecs were computed from the frequency and time characteristics of original and reconstructed signals. Objective parameters used for the analysis are segmental signal-to-noise ratio (SNRseg), log likelihood ratio (LLR), weighted spectral slope (WSS), and perceptual evaluation speech quality (PESQ). Values of these parameters are computed using MATLAB routines [7, 11].

Values of objective quality measures of CELP and CENLP codecs for different Indian dialects are listed in Table 4. Figure 7 shows the change in average values of quality measures among the above-mentioned codecs. Experimental results show that PESQ and SNRseg values are very good and high for CENLP. These measures give the enhancement in the quality of reconstructed signal with the use of proposed code-excited nonlinear prediction speech coder. LLR and MSE show low score for the proposed system. The agreement between the spectral magnitudes of original and decoded speech is indicated by LLR and its low value corresponds to close agreement.

6 Conclusion

In this modern era, the efficient transmission of audio and video signals at low bitrate has a greater requirement. CELP 4.8 is a good quality linear prediction speech coder with moderate bitrate and delay. In this work, a new speech codec is proposed by implementing nonlinear prediction in CELP coder to improve the quality by maintaining the bitrate and low delay. The proposed codec CENLP is validated through the subjective and objective quality analysis procedures. Processing delay is also measured for both CELP and CENLP algorithms. Both the algorithms have almost the same delay because of equal number of coefficients. Experimental results clearly say that, in CENLP there is an increase in the prediction gain and thereby an increase in quality that approaches toll quality. Further improvement in the speech quality can be achieved by doing codebook optimization in CENLP.

References

1. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63:561–580
2. Schroeder M, Atal B (1985) Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In: IEEE international conference on acoustics, speech, and signal processing, vol 10, pp 937–940
3. Chen J-H, Cox RV, Lin YC, Jayant N, Melchner MJ (1992) A low-delay CELP coder for the CCITT 16 Kb/s speech coding standard. IEEE J Sel Areas Commun 10(5):830–849
4. McCree A, Truong K, George EB, Barnwell TP, Viswanathan V (1996) A 2.4 Kb/s MELP coder candidate for U. S. Federal Standard. In: IEEE international conference on acoustics, speech and signal processing, vol 1, pp 200–203
5. ITU-I Rec. G.729 (1996) Coding of speech at 8 kbps using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)
6. Despotović V, Perić Z (2013) Design of nonlinear predictors for adaptive predictive coding of speech signals. In: 2013 21st Telecommunications forum (TELFOR), Serbia, Belgrade, 26–28 Nov 2013
7. Despotovic V, Goertz N, Peric Z (2012) Nonlinear long-term prediction of speech based on truncated Volterra series. IEEE Trans Audio Speech Lang Process 20(3):1069–1073
8. Chu WC (2003) Speech coding algorithms: foundation and evolution of standardized coders. Wiley

Table 4 Objective quality measures of CELP and CENLP codecs

	Objective measures											
Language	CELP						CENLP					
	LLR	SNRseg	WSS	PSNR	PESQ	MSE	LLR	SNRseg	WSS	PSNR	PESQ	MSE
Malayalam	0.65	1.11	44.75	125.67	2.27	0.028	0.58	1.32	60.43	132.45	3.25	0.013
English	0.67	1.04	62.83	117.87	2.42	0.044	0.58	1.01	80.67	125.32	3.12	0.0.22
Tamil	0.55	1.21	68.34	132.55	2.45	0.034	0.57	1.49	89.32	139.27	2.89	0.012
Hindi	0.69	1.15	61.12	131.41	2.41	0.051	0.61	1.43	68.77	139.23	3.21	0.041
Kids	0.53	1.23	60.32	127.45	2.97	0.052	0.44	1.57	76.54	131.66	3.31	0.040
Average	0.62	1.15	59.47	126.99	2.50	0.042	0.56	1.36	75.15	133.59	3.16	0.026

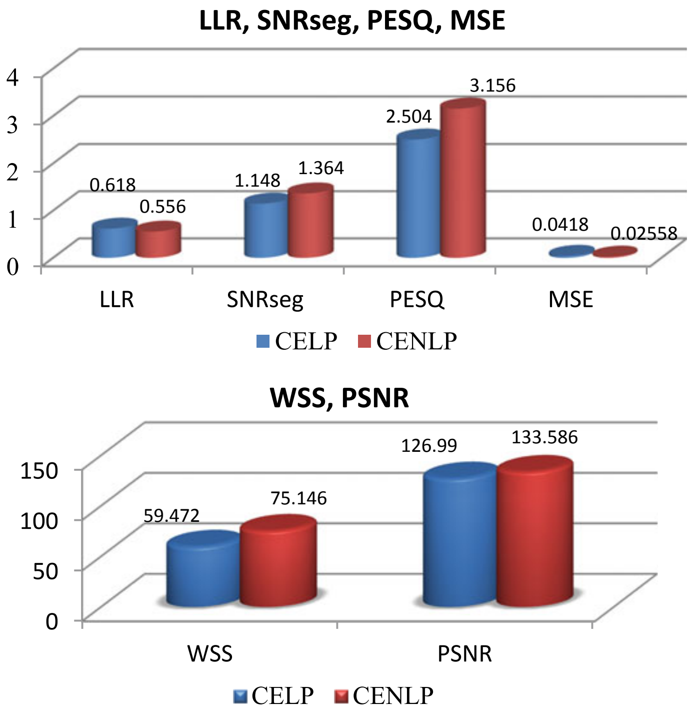


Fig. 7 Average objective quality measures of CELP and CENLP: LLR, SNRseg, PESQ, MSE, WSS and PSNR

9. Despotovic V, Goertz N, Peric Z (2012) Low-order Volterra longterm predictors. In: Proceedings 10th ITG symposium on speech communication, Braunschweig, Germany, Sept 2012, pp 26–28
10. Despotović V, Görtz N, Perić Z (2012) Improved non-linear long-term predictors based on Volterra filters. Vienna University of Technology, Institute of Telecommunications, Gußhausstr. 25–29, 1040 Vienna
11. Anselam AS, Pillai SS (2017) Optimization of code excited linear prediction speech coder with PSVQ-genetic codebook. In: 2017 international conference on wireless communications, signal processing and networking (WiSPNET)