ISSN 2277-3061

Performance Analysis of Advanced Hybrid Speech Coding Techniques in Time domain, Spectral domain and Perceptual domain

[1]Eslam Samy El-Mokadem, [2]Mohamed M. fouad, [3]Talaat A. Elgarf

[1]Communications Engineering Department, Higher Technological Institute, 10th of Ramadan city, Cairo, Egypt

[2]Communications Engineering Department, Faculty of Engineering- Zagazig University, Cairo, Egypt

[3]Communications Engineering Department, Higher Technological Institute, 10th of Ramadan city, Cairo, Egypt

ABSTRACT

Speech coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media and decoding the signal with the best possible perceptual Quality. The speech transmission in wireless networks is associated with the reduction of extra information present in signal in such a way to preserve the quality and intelligibility of speech. It is known that the lower the bit rate the lesser the quality of the reconstructed speech however there is a constant quest to achieve a better speech quality at lower bit-rates.

This paper presents performance analysis for the quality of advanced hybrid speech coding techniques in Time domain, Spectral domain and perceptual domain. These analyses are implemented on three different algorithms of advanced hybrid speech coding techniques such as CELP, G729 Annex A, G723.1 to assess the quality performance for English female speaker, English male speaker and Arabic female speaker by using Mat lab simulation program. Our evaluation criterion implemented includes the following tests: Signal to Noise Ratio (SNR), Segmental Signal to Noise Ratio (SNRseg), The Log-Likelihood Ratio (LLR), The Weighted Spectral Slope (WSS), Absolute Error, Perceptual Evaluation of Speech Quality (PESQ), Rating of speech distortion, rating of background noise and the predicted rating of overall quality.

Keywords

Speech coding; CELP; CSA-CELP; G729; Speech quality and Analytical evaluation

Council for Innovative Research

Peer Review Research Publishing System

Journal: INTERNATION JOURNAL OF COMPUTERS AND TECHNOLOGY

Vol. 13, No. 10

editorijctonline@gmail.com

www.ijctonline.com, www.cirworld.com

1. Introduction

CELP coder is widely used for mobile communication speech coding as a generic algorithm for implementing highly efficient and high-quality speech coding. Many standardized codecs are based on it. G729 and G723.1 are ITU standard speech codec based on CELP coder. G729 coder also called Conjugate Structure Algebraic CELP (CS-ACELP) coder. G723.1 also called Multi-pulse Maximum Likelihood Quantization (MP-MLQ). Both CELP, CS-ACELP and MP-MLQ encode speech in frames using linear predictive analysis by synthesis coding. This paper is organized as follows. In section 2, CELP speech coder is introduced. In section 3, ITU-T G.723.1 speech coder is introduced. In section 4, ITU-T G.729.1 speech coder is introduced. In section 5, various objective evaluation measures have been touched upon. In section 6, we describe MATLAB simulation for Objective Speech Quality Measures for the proposed coders. Performance evaluation of proposed coders is computed and demonstrated using set of tables and set of graphs. Finally the concluding remarks are given in section 7.

2. CELP Speech Coder

CELP coder provides the bridge among waveform coders and vocoders as it presents compression of speech comparable to medium bit rate waveform coders[1]. CELP algorithm is used to find the best code word characterizing the excitation signal for each 30 ms speech frame. This code word is found by applying each code word as an excitation for the CELP synthesizer. CELP is one of the most efficient speech coding algorithms where the speech is compressed with rate of 4.8 kbps by preserving quality of speech[1] .The synthesized speech signal is subsequently compared with the input speech signal and a difference signal is calculated. This difference signal is weighted by a perceptual weighting filter. As a result, the error signal e(n) is obtained from perceptual weighting filter [2]. That code word which ensures the lowest power of the error signal e(n) is selected as the best code word characterizing the frame. The characteristics of the formant weighting filter were chosen to ensure the best subjective human perception of the synthesized speech signal. The harmonic noise weighting filter controls the amount of error in the harmonics of the speech signal [2].

3. ITU-T G.723.1 Speech Coder

ITU-T G.723.1, the standard for multimedia communication speech coders, has two modes with bit rates of 5.3 and 6.3 kbit/s. The coder is based on the principles of linear prediction analysis-by-synthesis coding and attempts to minimize a perceptually weighted error signal [4]. The encoder operates on blocks (30 ms frame) of 240 samples each. Each frame is first divided into four sub frames of 60 samples each. In addition, there is a look-ahead of 7.5 ms, so the coder has a 37.5 ms total algorithmic delay. For every 60-sample sub frame, a set of tenth order LPC coefficients is computed. The LPC set of the last sub frame is converted to LSP parameters, and the LSP set is divided into 3 sub-vectors. The quantization is performed using a predictive split vector quantizer (PSVQ).The unquantized LPC coefficients are used to construct the short-term perceptual weighting filter, which is used to filter the entire frame speech and to obtain the perceptually weighted speech signal. For every two sub frames, the open-loop pitch lag is computed using the weighted speech signal. Every sub frame speech signal is then encoded by the ACB and FCB search procedures. The ACB search is performed using a fifth-order pitch predictor to obtain the closed-loop pitch and gains. Finally, the stochastic excitation pulses are approximated by MP-MLQ excitation for high bit rate (6.3 kbit/s), and ACELP for low bit rate (5.3 kbit/s) [5].

4. ITU-T G.729– ANNEX A Speech Coder

The general description of the coding/decoding algorithm is similar to ITU G729.

The G.729–ANNEX A is like G729 codec which is based on Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP). The coder operates on a speech frame (block) of 10 ms, which is equivalent to 80 samples at the sampling rate of 8000 Hz[6].Each block of 10 ms is first divided into two sub frames of 40 samples each. There is a 5 ms look-ahead for linear prediction (LP) analysis, resulting in a total 15 ms algorithmic delay For every 10 ms frame, the speech signal is analyzed to extract the parameters of the Code-Excited Linear-Prediction (CELP) coding model. A set of tenth order LPC coefficients are computed using the Levinson-Durbin algorithm. The LPC coefficients for the second sub frame are converted to LSP coefficients and are quantized using a predictive two stage vector quantizer. The unquantized LPC coefficients are used to construct the short term perceptual weighting filter. After computing the weighted speech signal, an open-loop pitch lag is estimated once per 10 ms frame based on the perceptually weighted speech signal. Next, the ACB and FCB are searched to obtain optimum excitation code vectors. ACB search is performed using a first-order pitch predictor, and a fractional pitch lag with one-third the sample resolution. In the FCB search, the stochastic excitation pulses are modeled using algebraic codebooks with four pulses [5].

The major algorithmic differences between G.729– ANNEX A and G729 are summarized below:

- The perceptual weighting filter uses the quantized LP filter parameters that are given by W(z) = Â (z)/ Â(z/ γ) with a fixed value of γ = 0.75.
- Open-loop pitch analysis is simplified by using decimation while computing the correlations of the weighted speech.
- Computation of the impulse response of the weighted synthesis filter W(z)/Â(z) computation of the target signal, and updating the filter states are simplified since W(z)/ Â(z) is reduced to 1/Â (z / γ).
- The search of the adaptive codebook is simplified. The search maximizes the correlation between the past excitation and the backward filtered target signal (the energy of filtered past excitation is not considered).

- The search of the fixed algebraic codebook is simplified. Instead of the nested-loop focused search, an iterative depth-first tree search approach is used.
- At the decoder, the harmonic post filter is simplified by using only integer delays.

This annex describes the changes to the full implementation which have been made in order to reduce the codec algorithmic complexity.

5. Objective Speech Quality Measures

The speech quality of a coding system can be linked to the perceived difference between the output of a system under test and a known reference signal. These differences are sometimes referred to as impairments. In evaluating the quality of a system, different types of Objective analysis have been carried out. This includes calculation of different parameters like [7]:

- ***Absolute error (ABSErr)***

The process of ABSErr computation is carried out by summing up the error values of each sample [8]. If $s(n)$ and $\hat{s}(n)$ represents the original speech signal and the synthesize speech signal respectively, then the error signal *e(n)* can be written as [7]:

$$e(n) = s(n) - \hat{s}(n) \quad \textbf{(1)}$$

Then, ABSErr can be given by:

$$ABSErr = \sum_{n} e(n) \quad \textbf{(2)}$$

- ***Percentage Error***

  Percentage error is calculated using the following formula:

$$\%Error = \frac{e(n)}{n} \quad \textbf{(3)}$$

- ***Mean Squared Error (MSE)***

  The mean squared error (MSE) defined as:

$$MSE = \frac{[e(n)]^2}{n} \quad \textbf{(4)}$$

- ***Root Mean Squared Error (RMSE)***

  RMSE is calculated as:

$$RMSE = \sqrt{MSE} \quad \textbf{(5)}$$

- ***Signal to Noise Ratio (SNR)***

  A widely used objective measure of Speech quality is the SNR. It is the ratio of the average energy in the original speech waveform to the average energy in the error signal. SNR represent the distortion introduced by the coding algorithm [8]. SNR can be calculated as follows:

$$SNR = 10\log_{10}\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N}\{x(n)-\hat{x}(n)\}^2} \quad [dB] \quad \textbf{(6)}$$

  where x (n) is the original speech, $\hat{x}(n)$ is the synthesized speech, and (N) the number of samples.

- ***Segmental Signal to noise ratio (SSNR)***

  SSNR is an improved of classical SNR, whereby the SNR measured over a quasi-stationary interval of 15–30 ms (Frames) and the individual SNR measures are averaged. SSNR makes distinction between errors that occur in high-energy regions and those in the low energy regions, where any errors will have a greater perceptual effect [9]. SSNR can be calculated as follows:

$$SNR_{seg} \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1}\{x(n)-\hat{x}(n)\}^2} \quad \textbf{(7)}$$

  Where L is the frame length (number of samples), and M the number of frames in the signal (N = ML).

- ***The Log-Likelihood Ratio (LLR)***

  It is a distance measure that can be directly calculated from the LPC vector of the clean and distorted speech .LLR measure can be calculated as follows:

$$d_{LLR}\ (a_d, a_c)\ = \log\ (\frac{a_d\ R_c\ a_d^T}{a_c\ R_c\ a_c^T}) \qquad \textbf{(8)}$$

  Where ($\mathbf{a}_c$) is the LPC vector for the original speech, ($\mathbf{a}_d$) is the LPC vector for the synthesized speech, ($\mathbf{a}^T$) is the transpose of$\mathbf{a}$, and ($\mathbf{R}_c$) is the auto-correlation matrix for the clean speech.

- ***The Weighted Spectral Slope (WSS)***

  It is a direct spectral distance measure. It is based on comparison of smoothed spectra from the original and synthesized speech samples. The smoothed spectra can be obtained from either LP analysis. WSS can be defined as follows:

$$d_{wss}=\frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K}W(j,m)(S_c(j,m)-S_d(j,m))^{\ 2}}{\sum_{j=1}^{k}W(j,m)} \qquad \textbf{(9)}$$

  Where K is the number of bands, M is the total number of frames, and Sc(j, m) andSd(j , m) are the spectral slopes (typically the spectral differences between neighboring bands) of the j th band in the m th frame for clean and distorted speech, respectively. W (j, m) are weights.

- ***Perceptual Evaluation of Speech Quality (PESQ)***

  It is an international standard (ITU-T recommendation P.862) for estimating the Mean Opinion Score (MOS) from both the original signal and its degraded signal. PESQ record the difference between the original signal and the synthesized signal and derive a score from 0 to 5 where 5 are the best.

  PESQ score is computed as a linear combination of the average disturbance value Diand and the average asymmetrical disturbance values Ai and   as follows [10].

  PESQ =a_0 + a_1Diand + a_2Aiand **(10)**

  Where a_0 = 4.5, a_1= 0.1 and a_2 = - 0.030

6. MATLAB Simulation Results

To compare the performance of each implemented codec algorithms, a simulation using MATLAB program is carried out with quality measurements to measure the quality performance for different algorithms in Time domain , Spectral domain and perceptual domain. These measures are applied on three algorithms of hybrid speech coders to test the quality performance of each algorithm for English female speaker, English male speaker, Arabic female speaker and Arabic male speaker. The objective performance evaluation of speech files includes calculation of parameters like Absolute Error, Mean Square Error, Signal to Noise Ratio, segmental Signal to Noise Ratio, Perceptual Evaluation of Speech Quality, The Log-Likelihood Ratio, Weighted Spectral Slope, rating of speech distortion, rating of background noise and the predicted rating of overall quality respectively.

6.1 Measuring the Quality performance of three algorithms for English female speaker using sound file (f1058.wav)

The wave file is used here for the purpose of this analysis, is (f1058.wav) for English female speech having 22630 samples. Equations utilized to calculate the above parameters are as inked in section V. MATLAB simulated mathematical results in Table 1 and graphical resulting plots are shown in Fig. 1, 2, 3. Results obtained by the objective analysis are found to be satisfactory as can be judged from figures cited at below [11].

Speech Reconstruction

The following graphs show the original signal, reconstructed signal, mixed signal between original and reconstructed and (SNR) for the three algorithms (CELP, G729-ANNEX A andG723.1 for English female speech file. The first part of each graph shows the original speech signal (blue) and the second part of each graph shows the reconstructed speech signal (red). The third part of each graph shows the mixed signal. The fourth part of each graph shows the curve for (SNR). The waveforms for CELP, G729-ANNEX A andG723.1 are shown in fig. 1, 2 and 3 respectively [11].

1. CELP

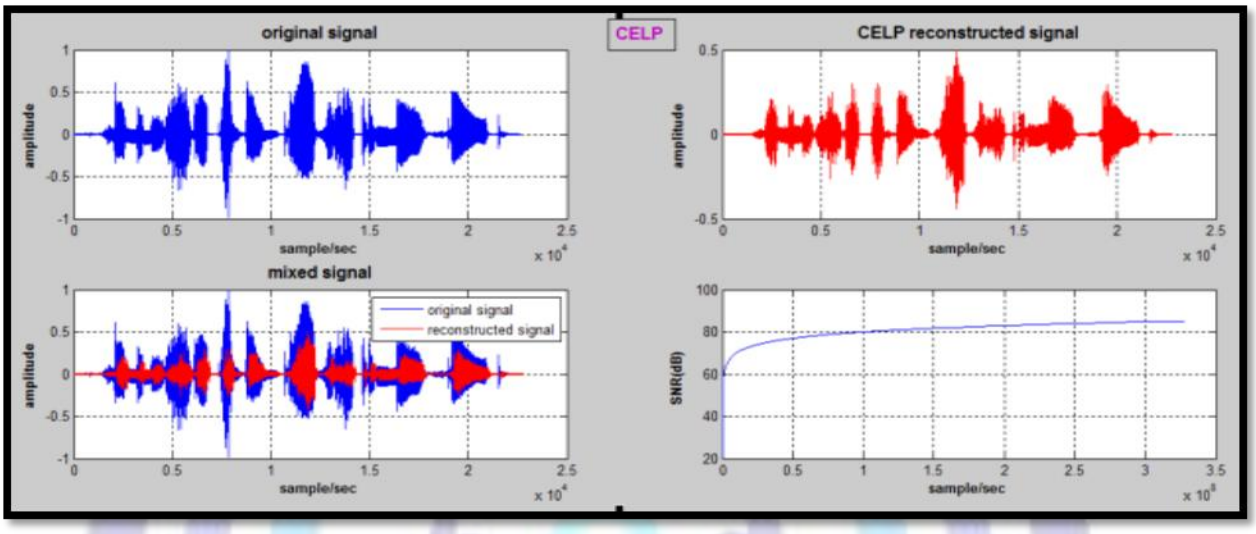


Fig 1: CELP Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of CELP [11].

2. G723.1

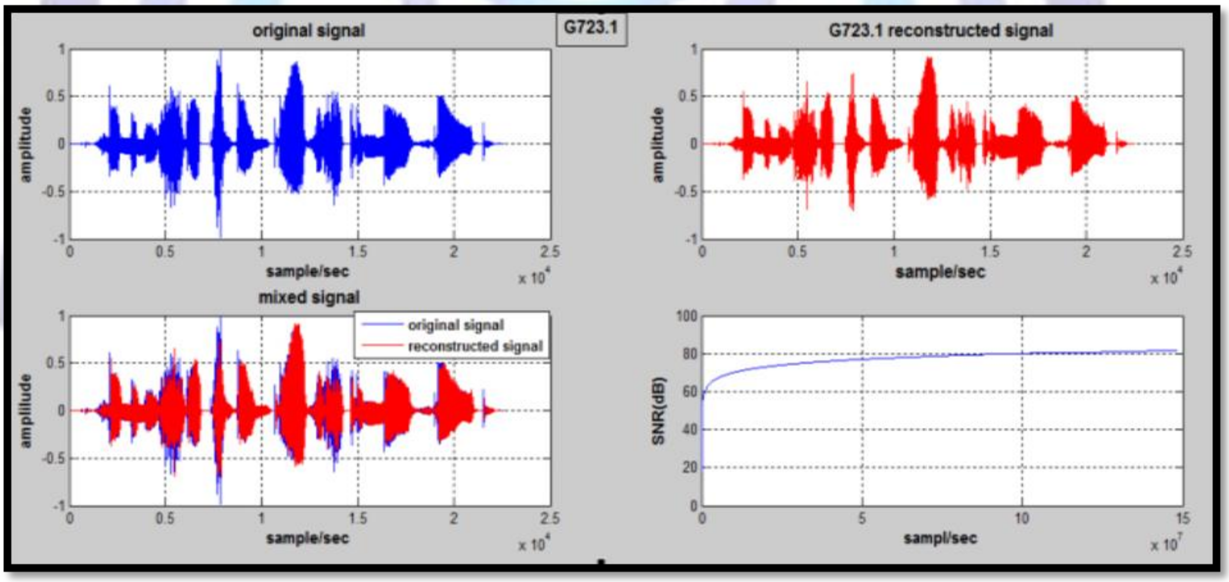


Fig 2: G723.1 Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G723.1 [11].

3. G729.1 ANNEX A

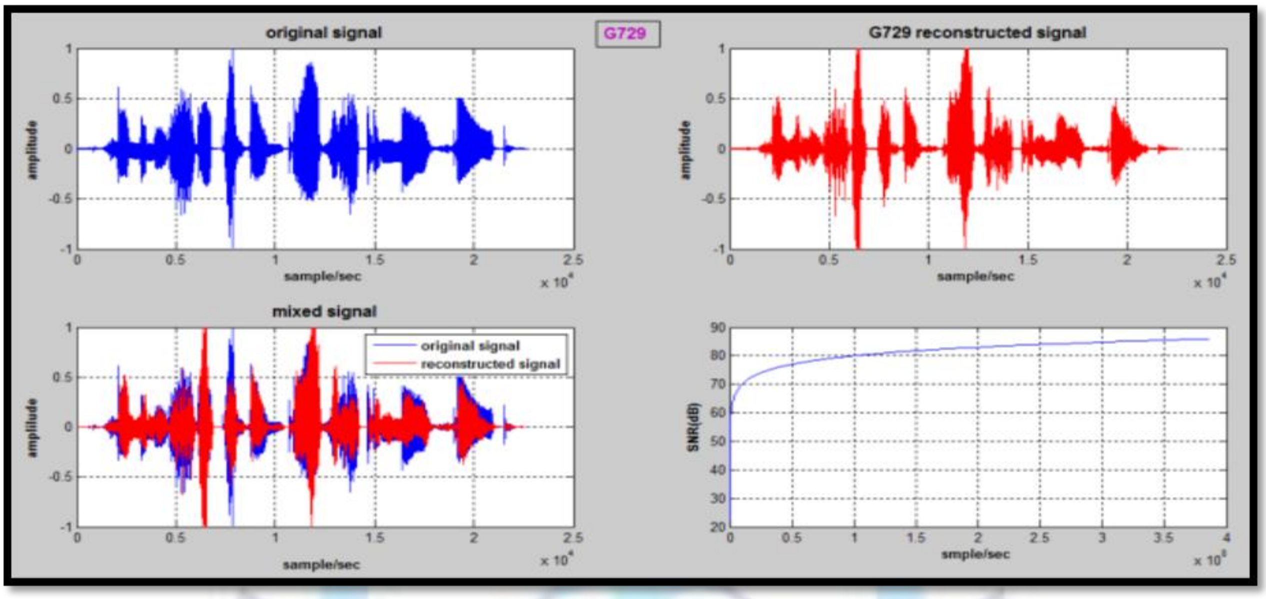
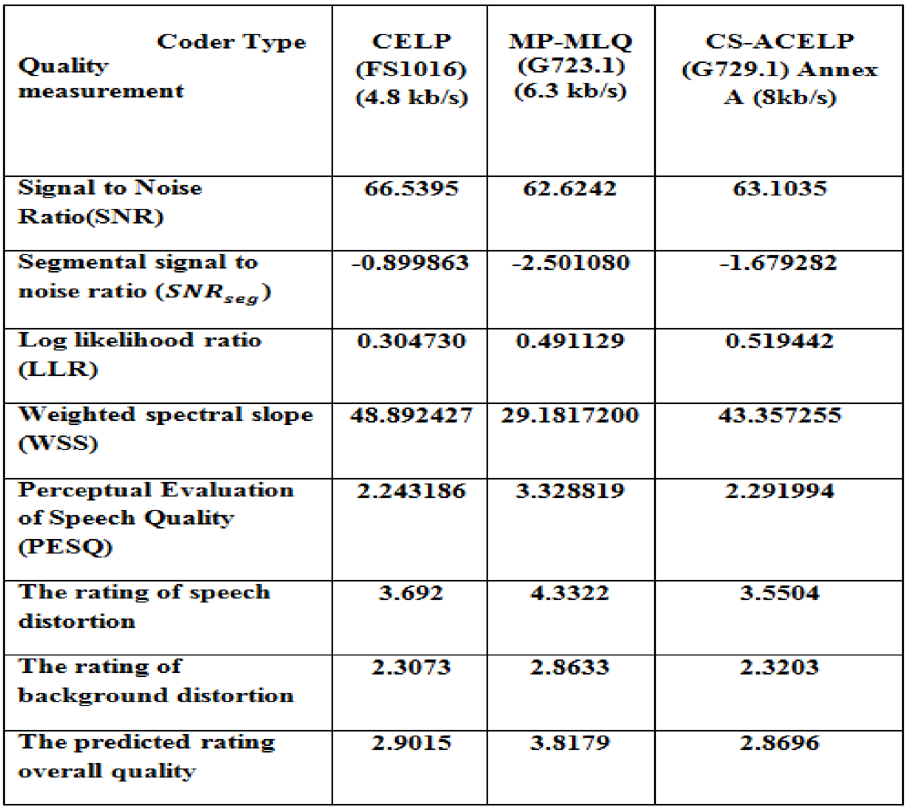


Fig 3: G729.1 ANNEX A Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G729.1 [11]

Table 1: Calculations of Speech Quality for English female speech file (f1058.wave) Using Different Coders



Coder Type Quality measurement	CELP (FS1016) (4.8 kb/s)	MP-MLQ (G723.1) (6.3 kb/s)	CS-ACELP (G729.1) Annex A (8kb/s)
Signal to Noise Ratio(SNR)	66.5395	62.6242	63.1035
Segmental signal to noise ratio (SNR_{seg})	-0.899863	-2.501080	-1.679282
Log likelihood ratio (LLR)	0.304730	0.491129	0.519442
Weighted spectral slope (WSS)	48.892427	29.1817200	43.357255
Perceptual Evaluation of Speech Quality (PESQ)	2.243186	3.328819	2.291994
The rating of speech distortion	3.692	4.3322	3.5504
The rating of background distortion	2.3073	2.8633	2.3203
The predicted rating overall quality	2.9015	3.8179	2.8696

6.2 Measuring the Quality performance of three algorithms for English male speaker using sound file (male.wav)

The wave file is used here for the purpose of this analysis, is (male.wav) for English male speech having 408226 samples. Equations utilized to calculate the above parameters are as inked in section V. MATLAB simulated mathematical results in Table 2 and graphical resulting plots are shown in Fig. 4, 5, 6. Results obtained by the objective analysis are found to be satisfactory as can be judged from figures cited at below.

1. CELP

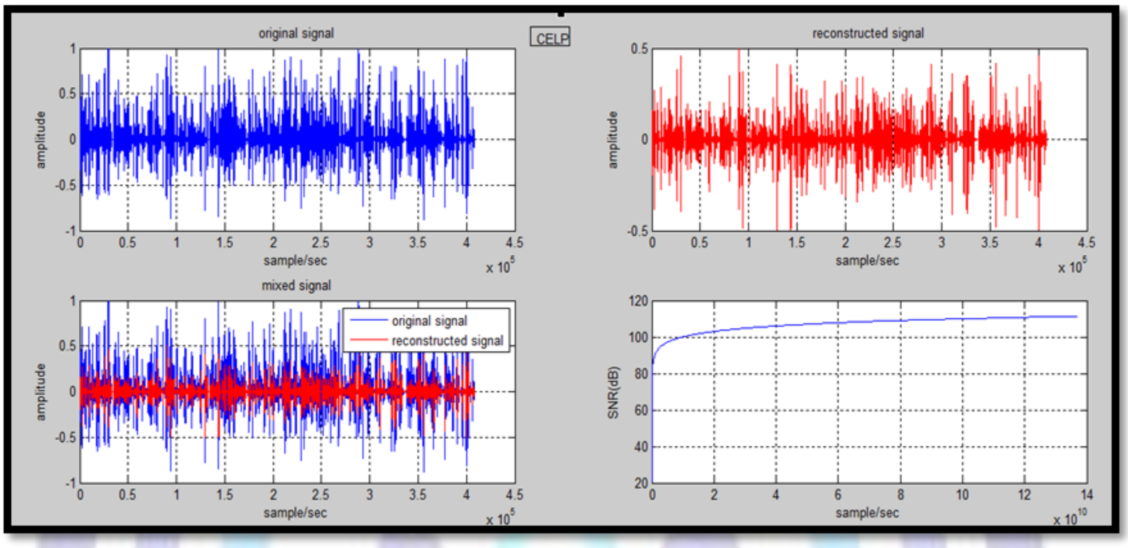


Fig 4: CELP Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of CELP

2. G723.1

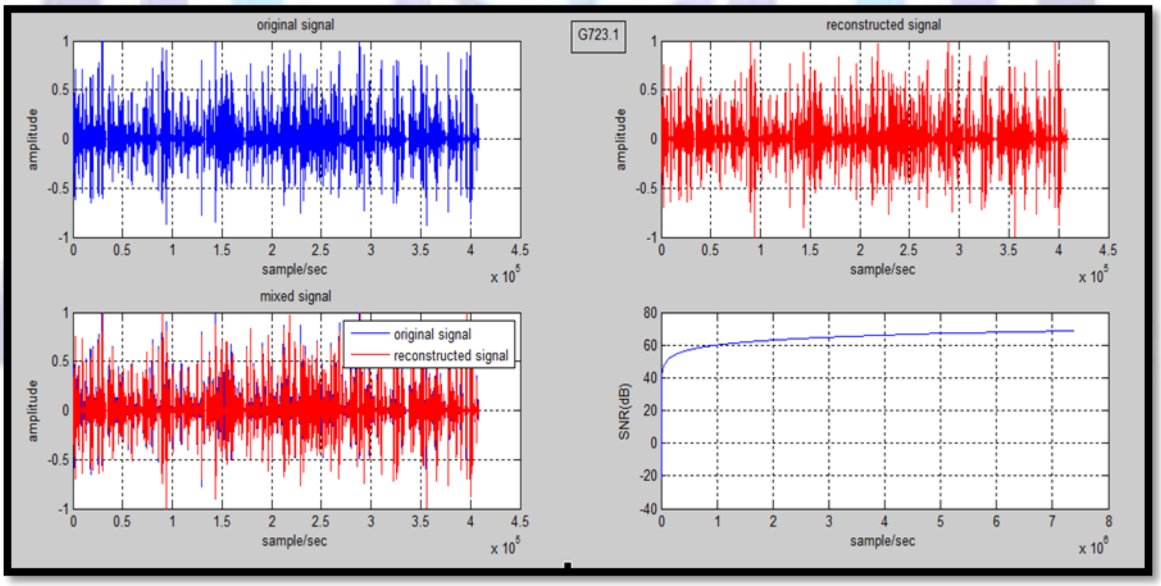


Fig 5: G723.1 Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G723.1

3. *G729.1 ANNEX A*

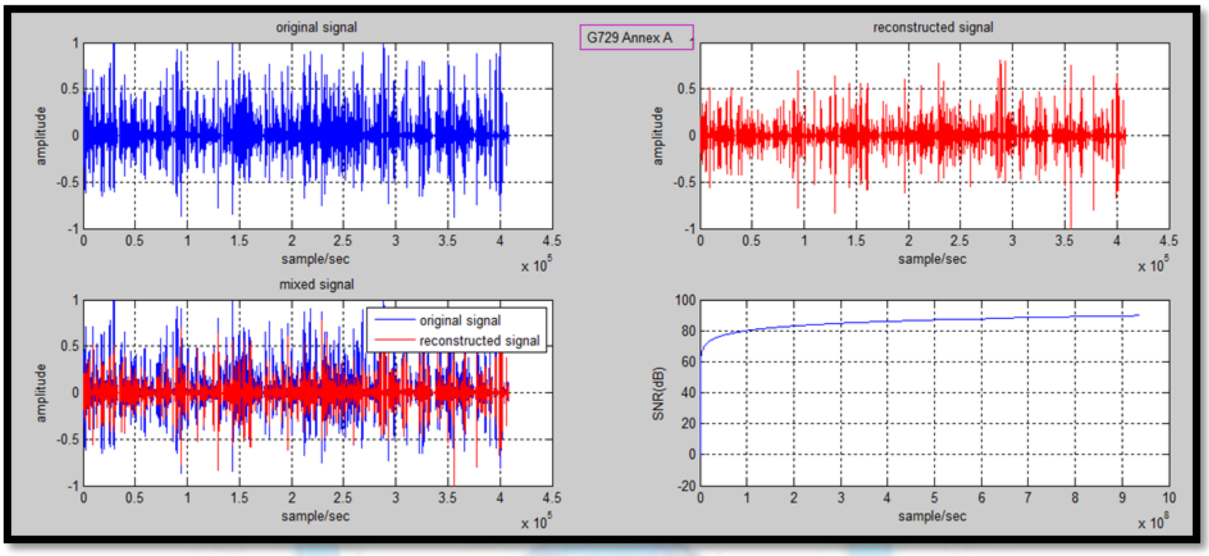


Fig 6: G729.1 ANNEX A Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G729.1

Table 2: Calculations of Speech Quality for English male speech file (male. wave) Using Different Coders

Coder Type Quality measurement	CELP (FS1016) (4.8 kb/s)	MP-MLQ (G723.1) (6.3 kb/s)	CS-ACELP (G729.1) Annex A (8kb/s)
Signal to Noise Ratio (SNR)	43.8907	47.2070	38.2390
Segmental signal to noise ratio (SNR_{seg})	-4.011256	-3.011256	-4.011256
Log likelihood ratio (LLR)	0.303427	0.303427	0.303427
Weighted spectral slope (WSS)	43.764452	37.764452	48.764452
Perceptual Evaluation of Speech Quality (PESQ)	2.829934	3.784602	2.558835
The rating of speech distortion	4.0933	4.6690	3.9299
The rating of background distortion	2.4276	2.8840	2.2981
The predicted rating overall quality	3.4104	3.1922	4.1789

6.3 Measuring the Quality performance of three algorithms for Arabic female speaker using sound file (Test.wav)

The wave file is used here for the purpose of this analysis, is (test.wav) for Arabic female speech having 58000 samples. Equations utilized to calculate the above parameters are as inked in section V. MATLAB simulated mathematical results in Table 3 and graphical resulting plots are shown in Fig. 7, 8, 9. Results obtained by the objective analysis are found to be satisfactory as can be judged from figures cited at below.

1. CELP

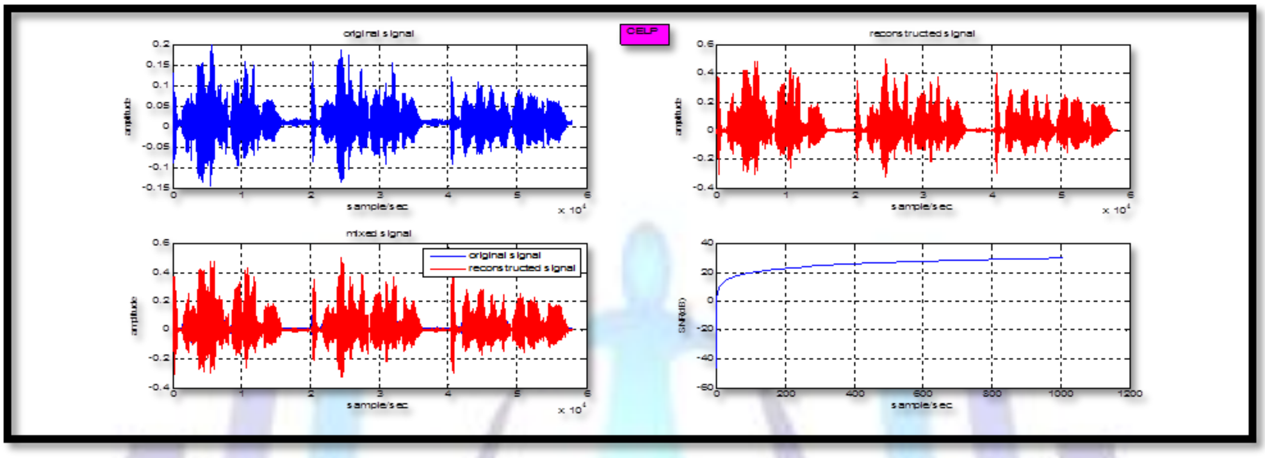


Fig 7: CELP Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of CELP

2. G723.1

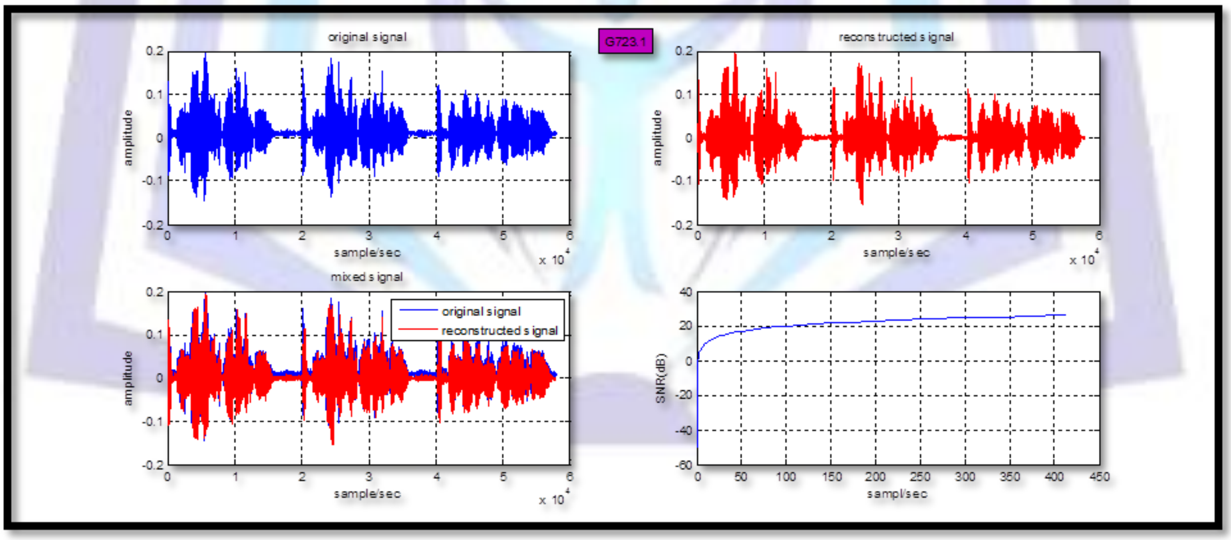


Figure 8: G723.1 Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G723.1.

3. *G729.1 ANNEX A*

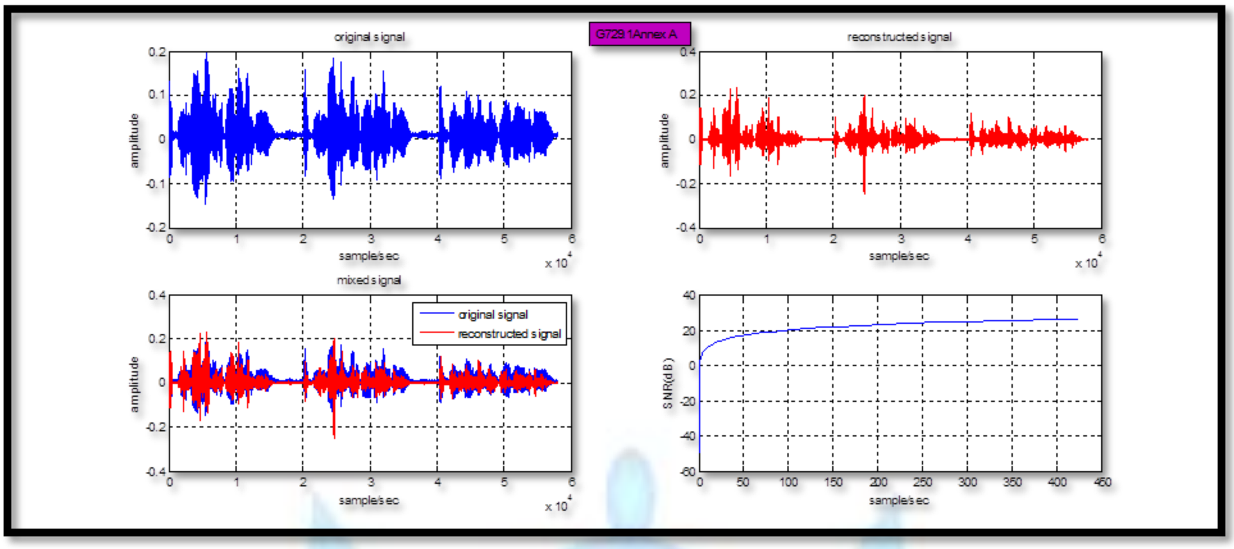


Fig 9: G729.1 ANNEX A Simulations: (a) Time Domain Input Speech, (b) Reconstructed Speech Signal, (c) Mixed Signal and (d) SNR of G729.1

Table 3: Calculations of Speech Quality for Arabic Female speech file (Test. wave) Using Different coders

Coder Type Quality measurement	CELP (FS1016) (4.8kb/s)	MP-MLQ (G723.1) (6.3kb/s)	CS-ACELP (G729.1) Annex A (8kb/s)
Signal to Noise Ratio (SNR)	12.6665	8.8228	9.5983
Segmental signal to noise ratio (SNR_{seg})	-1.309077	-2.771272	-1.23097
Log likelihood ratio (LLR)	0.782406	0.593841	0.752236
Weighted spectral slope (WSS)	56.536371	38.541926	52.416791
Perceptual Evaluation of Speech Quality (PESQ)	2.830551	3.507778	2.754189
The rating of speech distortion	3.4859	4.2503	3.4399
The rating of background distortion	2.5088	2.8663	2.4723
The predicted rating overall quality	3.0762	3.8439	3.0148

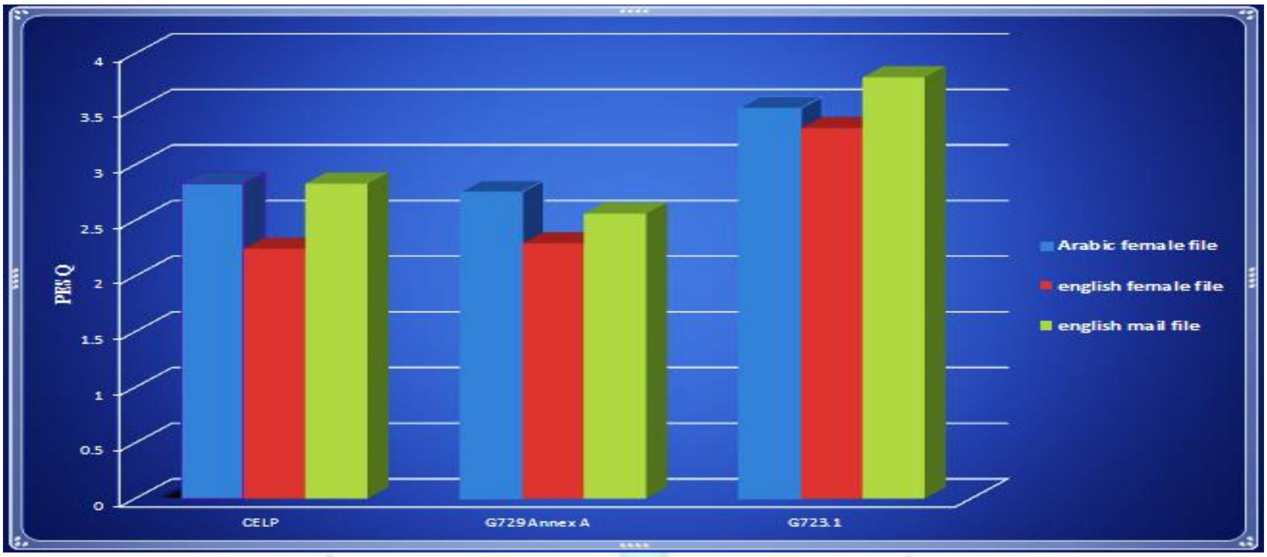


Figure 10: PESQ computation for three coders

7. Conclusion

This paper presents a performance analysis to assess the quality performance of advanced hybrid speech coding techniques in Time domain, Spectral domain and perceptual domain. evaluation criterion are implemented on three different algorithms of advanced hybrid speech coding techniques such as CELP, G729 Annex A, G723.1 to assess the quality performance for English female speaker, English male speaker and Arabic female speaker. by using Mat lab simulation program. Our evaluation criterion implemented includes the following tests: Signal to Noise Ratio (SNR), Segmental Signal to Noise Ratio (SNRseg), The Log-Likelihood Ratio (LLR), The Weighted Spectral Slope (WSS), Absolute Error, Perceptual Evaluation of Speech Quality (PESQ), Rating of speech distortion, rating of background noise and the predicted rating of overall quality. As can be seen from the obtained results and graphs, the quality of each codec is still good and can be heard but the analytical results proved that G723.1 is better than both of CELP and G.729– ANNEX A despite of G.729– ANNEX A has a bit rate higher than G723.1 and CELP. Also we can see that the quality performance of G729– ANNEX A coder is the lowest performance for English male speakers and Arabic female speaker, on the other side G.729 – ANNEX a slightly better performance than CELP coder but lower than G723.1 for English female speakers.

This means that the changes that have been applied to G729 by this annex in order to reduce the codec algorithmic complexity affected on the quality performance of this coder. It is observed generally that any increase of complexity of any algorithm will lead to an increase in delay time. In the future we may find a method to reduce the complexity of G.729– ANNEX A and G723.1 while maintaining the speech quality.

8. REFERENCES

[1] Nasir Saleem, Usman Khan, Imad Ali. "Implementation of Low Complexity CELP Coder and Performance Evaluation in terms of Speech Quality." International Journal of Computer Applications, September 2012.

[2] Eko Pryadi, Kuniwati Gandi, Herman Y. Kanalebe."SPEECHCOMPRESSION USING CELP SPEECH CODING TECHNIQUE IN GSM AMR." Wireless and Optical Communications Networks, WOCN '08. 5th IFIP International Conference on. pp: 1 - 4 , 2008.

[3] D.L .Neuhoff, "Design of a CELP Coder and Analysis of Various Quantization Techniques",EECS 651, Project Report University of Michigan, Source Coding Theory 2005.

[4] ITU-T Rec. G.723.1, Dual Rate Speech Coder for Multimedia Communications at 5.3 and 6.3 kbit/s, 1996.

[5] Rong-San Lin, Jia-Yu Wang and Jeng-Shyang Pan," AN EFFICIENT TRANSCODING SCHEME FOR G.729 AND G.723.1 SPEECH CODECS: INTEROPERABILITY OVER THE INTERNET" International Journal of Innovative Computing, Information and Control, July 2012.

[6] ITU-T Rec.G.729, Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code Excited Linear Prediction, 1996.

[7] Milind Tandel, Vandana Shah,Bhavina Patel. "Implementation of CELP CODER and to evaluate the performance in terms of bit rate, coding delay and quality of speech". IEEE international conference on Acoustic, Speech and Signal Processing, pp: 86-89, 2011.

[8] Eman Mohammed Mahmoud, Talaat A. Elgarf ,Abd El-halim Zekry, Ahmed Abd Elhafez "Implementation and evaluation of variable bit rate CELP CODER" , international conference on computer engineering and systems, November 2012.

[9] Hanzo, Lajos. “Voice and Audio Compression for Wireless Communications”, 557-558. John Wiley & Sons, Ltd, 2007

[10] Yi Hu, Philipos C. Loizou. "Evaluation of Objective Quality Measures for Speech Enhancement", IEEE Transactions on Audio, Speech and Language processing, Jan. 2008.

[11] Eslam Samy El-Mokadem, Mohamed M. Fouad, Talaat A. Elgarf," Evaluation Criterion to Assess the Quality Performance of Advanced Hybrid Speech Coding Techniques", IJCST Vol.4, Issue 3, July -September 2013.

Author' Biography with Photo

Eslam Samy El-mokadem is currently a M.S. student in Department of Communication at Zagazig university . He obtained his B.S. degree in 2007 from Higher Technological Institute (HTI) 10th of Ramadan City, Egypt. He is a teaching assistant, in the department of Electrical and computer engineering at Higher Technological Institute since 2008 until now. His research interests in speech coding techniques, communication systems, network security and wireless sensor networks.