Performance Evaluation of Speech CODECs against the Change in the Spoken Language and Accent

ABSTRACT

In all modern communications, speech coding plays a very vital role. It is a basic element in the efficient use of BW and QoS issues. Many standard bodies generate a series of speech coders with different rate, quality and delay combinations. Most of these CODECs have been built for 7 languages not including the Arabic or its accents. In this paper an extensive testing is done on the popular ITU G.723.1 and compared with the waveform G711, μ_law. The test is done for (English, Arabic and Cairo accent) from different speakers' gender. The assessment were developed using the PESQ algorithm and it has been proved that the speech quality will be slightly degraded when using other language or accent rather than English language. The main point is not only the slight degradation but also the less stability of the coder performance when working with languages other than English.

Keywords: performance, coders, accents, Arabic, Cairo accent, G.723.1, G.711, PESQ

I. Introduction

Speech coding is the process of digitizing the voice in a few number of bits, while maintaining good quality as possible. Speech coding has an important role in modern communication technologies, where quality and complexity have a direct impact on the market and the cost of the underlying products and services. It will remain in the center of attentions for years to come [1]. The general structure of the speech coding consists of: filter, sampler, ADC, source coder and finally the channel coder.

Wave Form coding is that type that simply works with the signal as amplitudes converting them to digital values, it is robust but very large BW required is for each signal for example for one way communication[2]:

- Sampling frequency = 8 KHz
- Number of bits per sample = 16 bits
- Bit Rate = 8 KHz * 16 = 128 Kbps

However, Linear Prediction Coding has an output rate that is much smaller than that of wave form coders it reaches about 2.4 Kbps with reduction of more than 50 times the wave form coders. In general the desirable properties of a speech coder [1] [3] can be summarized as follows:

- Low bit rate.
- High speech quality.
- Robustness against channel errors.
- Good performance in non speech signals.
- Low memory size and low complexity.
- Low delay.

In the previous work in this field of language and accent classification, the work of Hansen, Arslan [4] and Parry [5] has focused upon probabilistic decision as the origins of the speaker, while the work of Burnett, Parry [6] Itani, and Paulikas [7] handles the influence of language on LPC performance, they works only on Speexs which stands for "A Free Codec For Free Speech" and AMR "Adaptive Multi-Rate audio codec" Coders and with very small number of speech samples to get the results about the English, Arabic and Lithuanian.

The technique of the speech coder should be general enough to model different speakers (adult male, female... and children) with different language and accents; however this is not a trivial task. With most of the developments coming from English areas the problem arises that these advances may not be robust against other languages or accents. Problems will occur when the speaker uses a language or accent which contains phonemes that inherent to the speaker mother language and not contained in the English language [6].

II. Speech Quality Estimation

An essential requirement for all modern communication systems is the measure of the speech quality. This kind of measurement is either subjective or objective. The following figure (1) summarizes these categories. The oldest subjective way accepted internationally was the MOS (Mean Opinion Score). The MOS normally depends on asking people to grade the system by testing many calls however it is time consuming, expensive, and suffers lack of repeatability.

This in turn makes the objective methods more commonly used. The most widely used algorithm is the Perceptual Evaluation of Speech Quality PESQ. It involves comparison of reference and degraded speech signals to obtain a predicted listening only one way MOS score as in figure (2). It is standardized as ITU-T recommendation P.862 [8] [9].

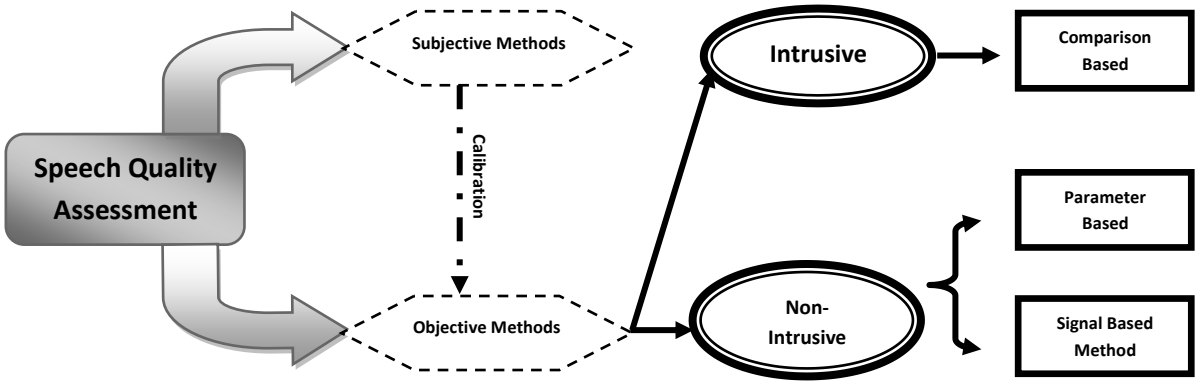

Fig. 1: Classification of speech quality assessment methods

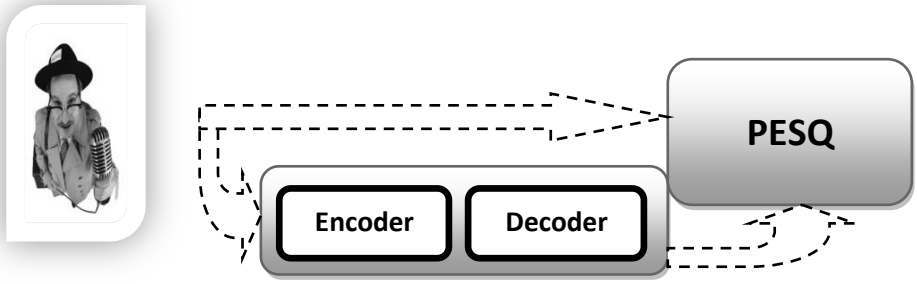

Fig. 2: PESQ Test

III. Sample of the Most Used CODECs.

A. The G.711 CODEC

One of the most widely used simple CODEC is the G.711 which is based on the Companding. Companding comes from Compression and Expanding to increase the Dynamic range. Voice is sampled and coded using Pulse Code Modulation and then logarithmically compressed using μ_law as equation (1). Thus the coded word is converted from 14 bit input to 8 bit output [10].

$$F(x) = sgn(x)\frac{\ln(1+\mu|x|)}{\ln(1+\mu)} \quad (1)$$

Where: x: the input signal magnitude and F (x): The output Signal magnitude.

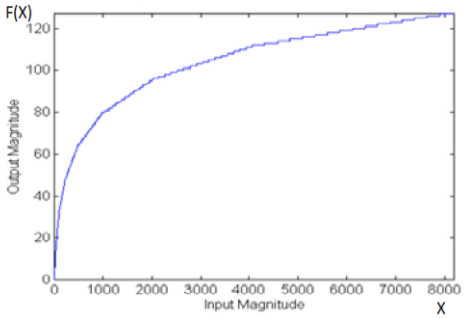

Fig. 3: μ_Law Compressor

The following figure shows the SIMULINK Model used to convert a saved speech file into G.711 and turn it back to Wave file again.

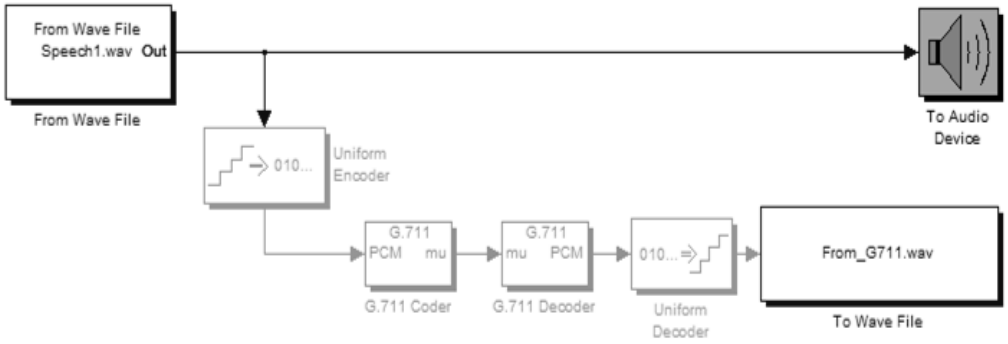

Fig. 4: G.711 Model

As the figure shows that it consists of 3 groups of SIMULINK blocks divided as the following:

- Group A: “From wave file”, “To wave File” and “To Audio Device”.
- Group B: The “Uniform Encoder” and the “Uniform Decoder”.
- Group C: The “G.711 Coder” and The “G.711 decoder”.

As stated before the main function for this model is to convert the previously saved speech wave file into G.711 and back to wave file. This will be done in many steps as follows:

- Open the pre saved wave file using the “From wave file” block.
- Prepare the speech to be encoded using G.711 CODEC which need a PCM input 14 bits per sample, and this will be done using the “Uniform Encoder” block.
- In the G.711 block we will select from either A_law or the μ_Law.
- The G.711 decoder and the Uniform decoder blocks operate to retrieve the original signal.
- The “To Audio Device” block used as immediate tester to see if the original signal file works or not.
- Finally the “To wave file” again used to save the decoded G.711 speech file to be further tested using the PESQ algorithm.

B. The G.723.1 CODEC

The ITU G.723.1 is one of the most popular LPC CODECs. It can work in either one of two modes, the MPC-MLQ (Multipulse LPC with Maximum Likelihood quantization) with a rate of 6.3 Kbps and the ACELP (Algebraic code excited linear prediction) with 5.3 Kbps. Each frame contains 240 samples for 30 ms duration [11] [12]. The following figure shows the Coder block diagram while the later one illustrates the Decoder.

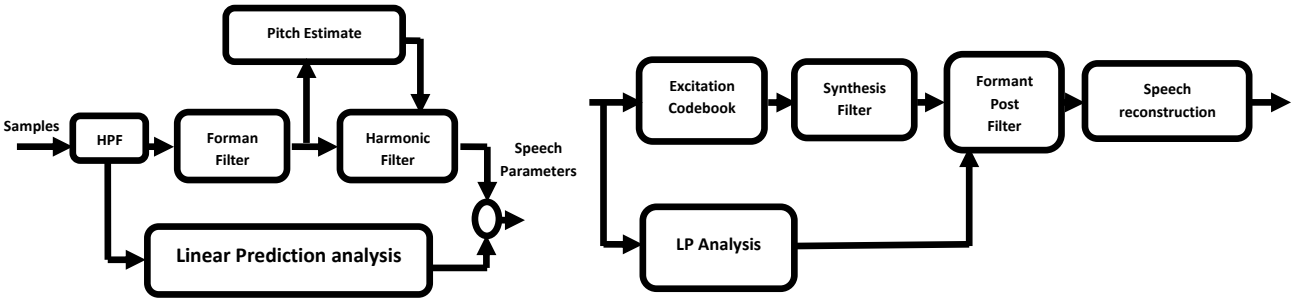

Fig. 5: G.723.1 Encoder (a) G.723.1 Decoder (b)

The coder operates on two time scales (i.e. frame scale 30 ms, 240 samples and sub-frame scale 7.5 ms, 60 samples). The coding process starts from the HPF, the high pass filter to remove the DC component. Then the Linear Prediction analysis is done on each sub-frame this creates 4 groups of LP coefficients. The output from the HPF also fed to the formant filter to extract the filter coefficients of the speech formants. The output from the formant filter is used to estimate the pitch of the sound. Both of the outputs from the formant filter and the pitch estimation will be used to get the harmonics associated with the speech using the harmonic weighting filter.

The decoder on the other hand must retrieve the original speech. It is also works on frame by frame basis. This will be done in four main steps. The speech parameters received from the coder will fed to both the LP analysis to reproduce the filters from the coefficients that have been sent after decoding them. The excitation code book is searched to generate the excitation signal. This signal then passed through the synthesis filter whose output is input to the formant post filter. The final gain scaling to maintain the energy level of the original signal will be adapted in the final step of the speech reconstruction.

IV. The Test Strategy and the Results Analysis.

A. The Test Strategy

The test strategy can be summarized in the following steps:

1. Save 10 voice files for English, Arabic and Cairo accent each 10 seconds long [with the same person].
2. Use the same files to generate the output of the G.711 PCM μ _Law and ACELP G.723.1 CODECs.
3. Use the objective speech quality measurement PESQ algorithm to estimate the quality.
4. Tabulate the result and conclude the effects.

B. Results analysis

Using 210 speech samples recorded at the same test environment, but with different talkers' age and sexuality. Every sample is passed through both G.711 and G.723.1 coder and decoder. The decoded speech undergoes the PESQ test to study many quality parameters.

The parameters under test are the equivalent PESQ MOS score, the signal to noise power ratio which is given by equation (2), the PRSD (The Predicted Rating of Speech Distortion), the PRBD (The Predicted Rating of Background Distortion), and the PROQ (The Predicted Rating of Overall Quality).

$$SNR = 10\log\left(\frac{\sum_n X[n]^2}{\sum_n (X[n]-Y[n])^2}\right) \quad (2)$$

Where: X[n]: the speech samples. And Y[n]: the noise samples [1].

The following tables illustrate samples of the PESQ output for two males and two females. For better estimation of the language or accent change we will use the standard deviation statistic value. The standard deviation shows how much dispersion exists from the mean value. Low standard deviations indicate the data values are very close to the average which means better signal stability and vice versa. Table (5) shows the standard deviation for the tabulated above samples.

Table 1: 1st Male PESQ MOS Score for G.723.1

PESQ_AR	PESQ_Cairo Accent	PESQ_EN
3.3583	3.3715	3.3395
3.1569	3.2888	3.4025
3.1704	3.335	3.3686
3.2962	3.2966	3.4295
3.3793	3.2924	3.3787
3.5436	3.2338	3.4694
3.321	3.346	3.4259
3.4499	3.215	3.4407
3.2482	3.2963	3.4376
3.2556	3.3306	3.3597

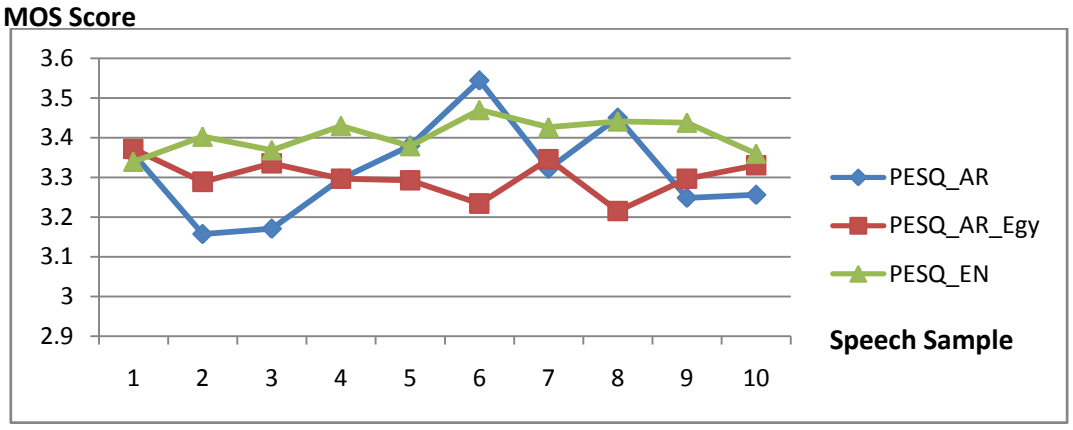

Fig. 6: PESQ MOS Score for G.723.1versus Sample number for 1st male

Table 2: 2nd Male PESQ MOS Score for G.723.1

PESQ_AR	PESQ_Cairo accent	PESQ_EN
3.0701	3.1736	3.1922
3.2893	3.1299	3.2439
3.1273	3.1238	3.1384
3.13	3.4298	3.1952
3.2102	3.1866	3.2838
3.1636	3.3172	3.1504
3.1423	3.3196	3.1751
3.2271	3.1721	3.2138
3.2493	3.2308	3.1941
3.1289	3.2484	3.2374

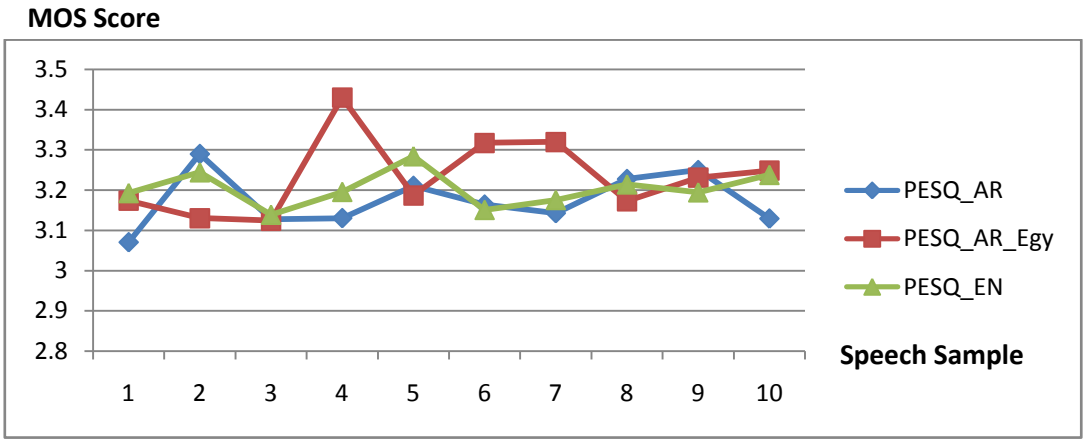

Fig. 7: PESQ MOS Score for G.723.1 versus Sample number for 2nd male

Table 3: 1st Female PESQ MOS Score for G.723.1

PESQ_AR	PESQ_Cairo Accent	PESQ_EN
2.98613	2.820762	2.85766
2.94946	2.849502	2.82312
2.86196	2.980833	2.81689
2.88401	2.916881	2.74867
2.84356	2.791334	2.77426
2.92785	2.761147	2.76747
2.90132	2.833235	2.70859
3.01674	2.906345	2.74006
2.81294	2.962018	2.70808
2.98811	2.901681	2.6692

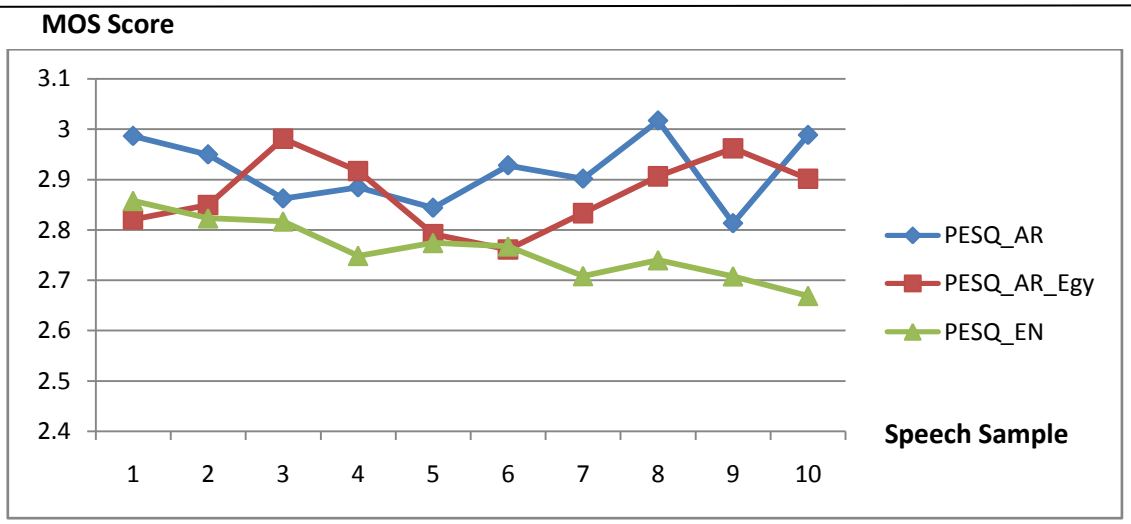

Fig. 8: PESQ MOS Score for G.723.1 versus Sample number for 1st Female

Table 4: 2nd Female PESQ MOS Score for G.723.1

PESQ_AR	PESQ_Cairo Accent	PESQ_EN
3.0278	3.0203	3.0372
3.0335	3.0268	3.0365
3.009	3.0366	3.0513
2.9747	2.9507	2.977
2.9799	3.072	3.0098
3.0319	2.8742	3.0067
2.9365	2.9937	2.9816
3.168	3.029	3.0376
3.0115	2.8831	2.9353
2.9642	3.1016	2.9639

Fig. 9: MOS versus Sample number for 2nd Female

Table 5: The Standard deviation values for the sample PESQ output

	1st Male	2nd Male	1st Female	2nd Female
Arabic	0.114308	0.064151	0.064453	0.059859
English	0.039858	0.041801	0.055773	0.036226
Cairo Accent	0.046046	0.092469	0.068885	0.071249

Table 6: 1st Male SNR for G.711 and G.723.1

SNR_AR_G.711	SNR_AR_G.723.1
2.499383	-0.126
2.499615	-0.100961
2.499491	-0.025654
2.499639	-0.065297
2.499727	-0.155991
2.499427	-0.251228
2.49946	-0.093799
2.499373	-0.141971
2.499597	-0.081256
2.499546	-0.076966

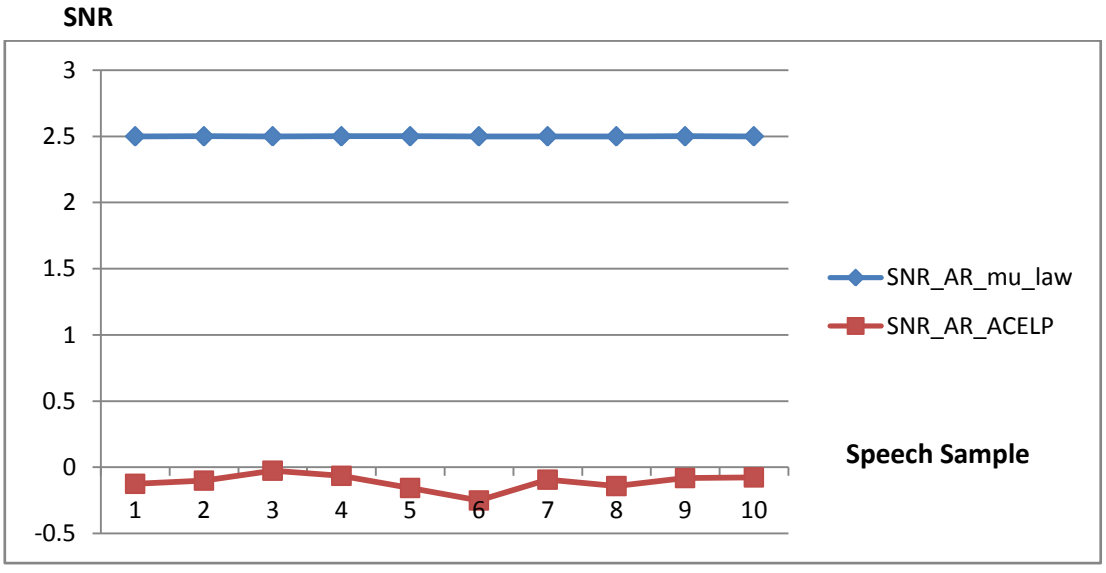

Fig. 10: 1st Male SNR for G.711 and G.723.1 Vs sample number

Table 7: 1st Male PROQ for G.711 and G.723.1

PROQ_AR_G.711	PROQ_AR_G.723.1
4.0983	2.0967
3.9075	2.3118
3.7019	1.4575
3.9241	1.6856
4.0961	2.1423
4.2035	2.3483
4.0237	2.0081
4.1182	2.2667
3.7657	1.7821
3.9371	1.7723

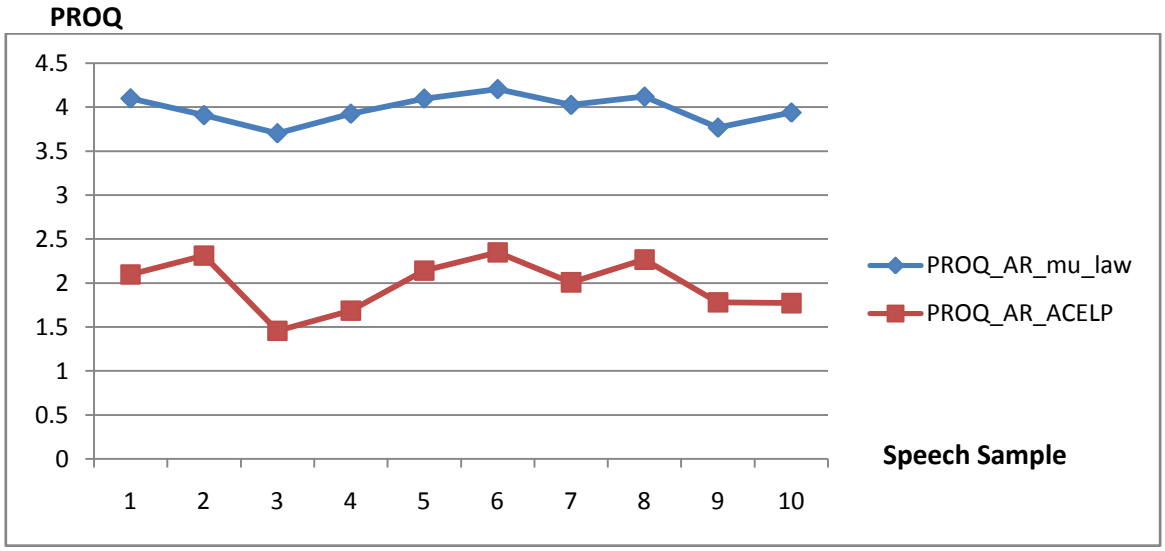

Fig. 11: 1st Male PROQ for G.711 and G.723.1 Vs sample number

From the above tables and curves it can be shown that:

1. The MOS of English recorded speech is very slightly higher for most cases than either Arabic or Cairo accented Arabic as you can see from tables 1, 2, 3 and 4.
2. The most important result is not the exact values of MOS but the instability that appears clearly in either Arabic or Cairo accented Arabic curves as shown in figures 7, 8, 9, and 10.
3. The above result is also confirmed by the lower standard deviation for English than either Arabic or Cairo accented Arabic, table 5.
4. G.723.1 shows better results of MOS scores for males rather that for females, tables 1, 2, 3, and 4.
5. The values of MOS is always better for G.711 than that of the G.723.1, however this will be in the expense of the band width.
6. The G.711 SNR is almost constant and higher than that of G.723.1. This is because of the SNR measure is suitable for the waveform coders rather than for the LPC. The wave form coders depends on the

number of bits per sample which is fixed for G.711, and hence the constant SNR value. In G.723.1 the SNR value always almost negative near to zero which means mathematically that the signal power less than the noise power. This is actually not true but the SNR only can compare the original signal with the received one which may greatly differ on shape with G.723.1 CODEC. For this reason we make use of PESQ algorithm which compares the formant properties of the speech rather than the signal error as SNR, figure 11 and table 6.

7. Finally it can be shown that from the last figure that the overall quality is rather better for G.711 over that of the G.723.1, as you can see in table 7 and figure 12.

V. The Conclusion.

It has been proved that the speech quality of the G.723.1 is better for English language rather than Arabic language or Cairo accent. G.723.1 gives slightly higher PESQ-MOS score for English over the Arabic or Cairo accent. It is also shown that the standard deviation of the PESQ-MOS score for English language is lower than that of Arabic or Cairo accent which means that it is more stable than them. Another observation was concluded that the PESQ- MOS scores for men are better than for women. This will lead us to develop an excitation new codebook to improve the performance of the G.723.1 CODEC with respect to Arabic language and Cairo accent which will be presented in another paper.

REFERENCES

1. WAI C. CHU," *Speech Coding Algorithms Foundation and Evolution of Standardized Coders*", 2003 by John Wiley & Sons, Inc.
2. Oppenheim and Schafer, "*Discrete-time signals and systems*", 1989.
3. Stephen E. Levinson, "*Mathematical Models for Speech Technology*", 2005 by John Wiley & Sons Ltd.
4. J.H.L. Hansen, L.M. Arslan, "*Foreign Accent Classification Using Source Generator Based Prosodic Features*", Proc. Int. Conf. Acoust. Speech Sign. Process. Detroit, pp. 836-839, 1995.
5. J.J. Parry, "*Accent Classification for Speech Coding*", Honours thesis, The University of Wollongong, 1995.
6. I. S. Burnett and J.J.Parry ," *On the Effects of Accent and Language on Low Rate Speech Coders* ".
7. Mohamad Itani, Šarūnas Paulikas, "Influence *of Language on CELP CODECS Performance*", ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2008, Vol.37, No.2.
8. ITU – T Recommendation P.862 -1993.
9. A.Bruce Carison, Paul B.Crilly, and Janet C.Rutledge," *Communication Systems: An Introduction to Signals and Noise in Electrical Communication*", McGraw Hill, 2002.
10. ITU – T Recommendation G.711 -2007.
11. ITU – T Recommendation G.723.1 -2009.
12. P. Kabal, ITU-T G.723.1 Speech Coder: "*A Matlab Implementation, Department of Electrical & Computer Engineering McGill University*", 2009.