The Effect of the Spoken Language on the Linear Prediction Vector Quantization Distortion for Linear Prediction Coders

Michael N. Micheal[1] ,Nagy W. Messiha[2], Hala A. Mansour[1], Hossam E. Mahmoud[1]

[1]Faculty of Engineering, Benha University, Cairo, Egypt, michaelnasief@yahoo.com

[2]Faculty of Electronic Engineering, Minofia University, Cairo, Egypt, dr.nagy_wadie@hotmail.com

ABSTRACT

Speech coding is the process of converting voice signal into digital form in a few bits as possible. The newly developed Code Excited Linear Prediction "CELP" coders is one major type of the parametric coders which combines between low data rates and good speech quality. Most of these coders have been built initially for 7 languages not included Arabic language or its dialects. It is known that the speech quality is directly proportional to the data rate but what is the effect of the change of the spoken language or accent? This paper is made to answer on three main questions. The first question is; what is the effect of the language or accents on CELP coders? Moreover what will happen if the speech is compressed more by lower data rate coder and at the same time the language is other than English? Finally if there is an effect, so what is the defective part in the coder? Extensive testing is done on 3 coders ITU G.711, ITU G.723.1 and 3GPP AMR with changing the spoken language. The outputs were compared with the ITU PESQ algorithm. It has been proved that the speech quality will be slightly degraded when using languages other than English. Also the quality is dramatically decreased as the compression ratio increased which will be rather lower and unstable for Arabic or Cairo accent. Finally it is found that the main problem was in the Linear Prediction vector quantization CodeBook and has been verified by the MFCC for English, Arabic and Cairo accent with LBG algorithm.

Keywords: performance, coders, accents, Arabic, Cairo accent, G.723.1, G.711, AMR, PESQ, MFCC, Mel, LBG, CodeBook.

1. Introduction

Speech coders classified in terms of operation as waveform, parametric and hybrid coders. The waveform coders is the simplest, most robust and highest data rates coders. It works with the signal amplitudes and the usual bandwidth for PCM for instance is 64 Kbps [1].

On the other hand parametric coders that see the speech as segments have parameters can reach data rates lower than 2 Kbps with acceptable speech quality. When designing any speech coders the general considerations or desirable properties are [2]

- Low data rate with high speech quality.
- Robustness against channel errors and different speakers and language.
- Good performance in non speech signals.
- Low memory size and low complexity and low delay.

The Code Excited Linear Prediction was first proposed in 1985. It is the basic principles of many other coders like ACELP "Algebraic" and CS-CELP "Conjugate Structure". Its basic working components [3]:

- Linear prediction model (LP).
- The use of the adaptive
- Fixed codebook to excite the LP model.

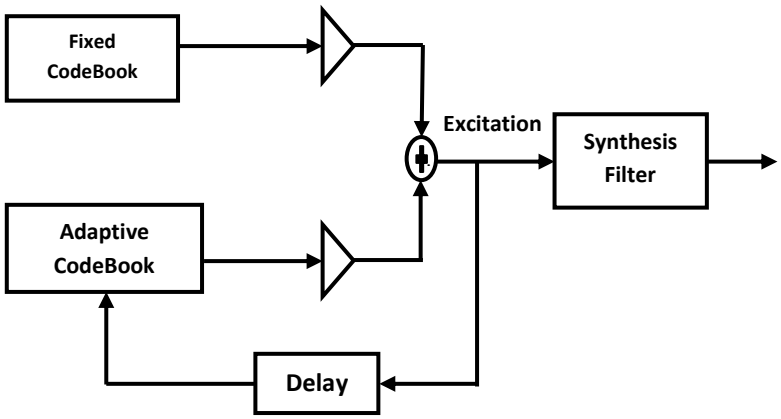

Fig. 1: The CELP model of speech synthesis.

The main principle of the CELP is the "Analysis by Synthesis". The idea of the AbS (Analysis by Synthesis) is to create many output signals using different choices of parameters. Each group of parameters will pass through PWF (perceptual weighting filter). The PWF is consisting of the formant filter and harmonic filter. The output from the PWF will be compared to the input speech that passed through the original perceptual weighting filter. The group of parameters that minimize the comparison error will be chosen and the indices associated to them will be sent.

In the previous work in this field of the effect of languages and accents on speech coders performance, 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], [8], [9] 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 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].

In this paper three steps were done. The purpose of the first and second parts is to investigate the problem existence of the effect of the spoken language on the LP coders. The third step is done to point out on the defective part of the coder and how to optimize it. Three coders were selected for the experimental tests made here; G.711 as high rate waveform coder, G.723.1 as CELP medium rate coder and AMR as medium rate coder. The first test is made with G.711 and G.723.1 to compare between the influences of the spoken language on the waveform and LP coders. In the second test the primary point for the selection of the AMR and the G.723.1 for the practical work is that, both lying on the same principle of ACELP, where AMR is 3GPP standards while the G.723.1 is ITU standard. It must be noted that the AMR used here is the full rate version at 12.2 Kbps while G.723.1 is used at 5.3 Kbps, then the effect of the bit rate reduction on the speech quality can be emphasized. In order to estimate the effect of the change of the spoken language or accent we selected English, Arabic and Cairo accent to compare the PESQ values and conclude the effects. The third test MFCC (Mel Frequency Cepstrum Coefficients) was selected to declare the difference of the formant distributions between the different languages. Finally the LBG (Linde Buzo Gray) algorithm was used to demonstrate the large amount of distortion when using Codebook not initially trained for Arabic language.

This paper organized as follows, the 2nd section is the proof of the problem existence (problem: the effect of the spoken language on the performance of the Linear Prediction Coders) while the third section gives the problem identification. Final section summarizes the conclusion.

2. The Proof of the Problem Existence.

In this section the problem existence of the effect of the spoken language on the LP Coders has been proven. This section introduced in three parts, part one gives a brief on the coders under test. The second part demonstrates the idea of the PESQ algorithm while the third section discusses the results of the practical work.

2.1. Coders under Test.

Three Coders used in the practical work will be introduced briefly here. These coders are ITU G.711 128 Kbps, ITU G.723.1 Multi-Pulse 5.3 Kbps and 3GPP AMR 4.75 to 12.2 Kbps.

2.1.1. The G.711 CODEC

The first coder developed by the ITU G Series is the G.711. it is the usual PCM with A_law or μ_law companding. Speech is sampled at 8000 samples per second then each sample is first coded into 14 bits word using uniform encoder then compressed to 8 bits. The companding using μ_law given by [10]:

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

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

A SIMULINK MATLAB model has been built to convert recorded speech files to and from G.711.

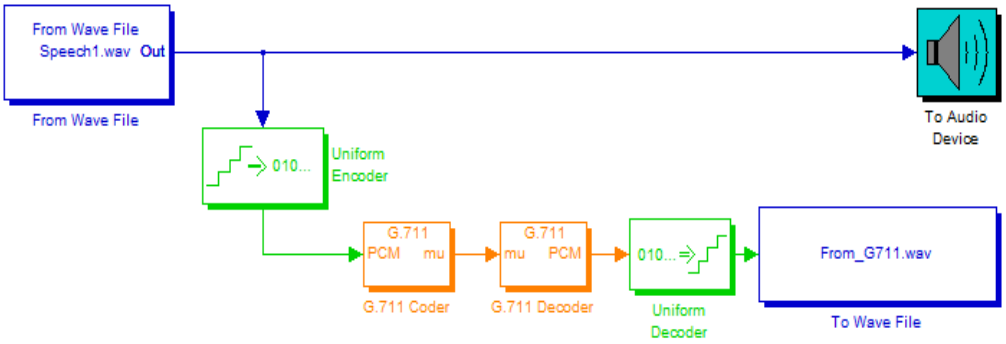

Fig. 2: G.711 Model

2.1.2. The G.723.1 CODEC

The ITU G.723.1 is a dual rate coder. It can work either in ACELP mode with 5.3 Kbps or MP (multi pulse) mode with 6.3 Kbps. The speech frame consists of 240 samples that were sampled at 8000 sample per second. The following figure shows the Coder block diagram while the later one illustrates the Decoder [11], [12], [13].

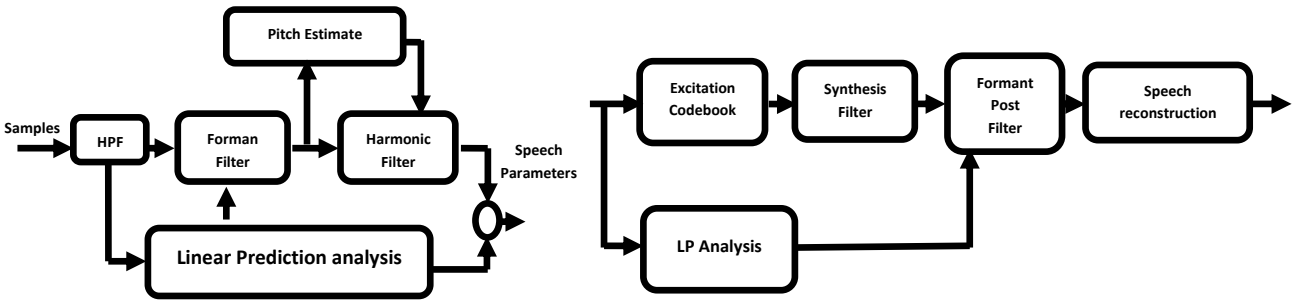

Fig. 3: G.723.1 Encoder (a) G.723.1 Decoder (b)

The input to the coder must first pass though high pass filter that removes any DC components associated with the input speech samples. The next step is the Linear Prediction Analysis. The output from the LP is fed to the Formant filter then to the open loop pitch estimation. Both the output from the pitch estimator and the formant filter are used to excite the fixed and adaptive codebooks. The final target signal will contains the quantized linear prediction coefficients and the gain, pitch, and pulse positions from the analysis by synthesis (not shown in the fig). The decoder essentially is the reverse operation.

2.1.3. Adaptive Multi-Rate "AMR".

Adaptive Multi-Rate (AMR) is a speech coding developed as the standard speech codec by 3GPP for the 2nd and the 3ed mobile generations [14]. The principle of the AMR CODEC is to use similar algorithms of many CODECs with different data rates. The communication network is responsible for the selection between these rates according to the channel robustness. For highly noise channel the lower rate will be used in order to increase the number of redundancy bits, so maintain the connection with acceptable speech quality. For good channel (i.e. free of errors) AMR full rate at 12.2 kbps will be used giving the best speech quality. The following table shows the different AMR rates [15], [16].

CODEC	Bit Rate
AMR Full Rate	12.2 kbps
AMR 10.2	10.2 Kbps
AMR 7.95	7.95 kbps
AMR 6.7	6.7 Kbps
AMR 5.9	5.9 Kbps
AMR 5.15	5.15 Kbps
AMR NB	4.75 kbps

Table 1: AMR Different Rates [16]

The AMR CODEC is an algebraic code excited linear prediction (ACELP) CODEC. It uses 10th order short term linear prediction filters and both algebraic and adaptive codebooks. Figure (4) shows the block diagram of the AMR coder.

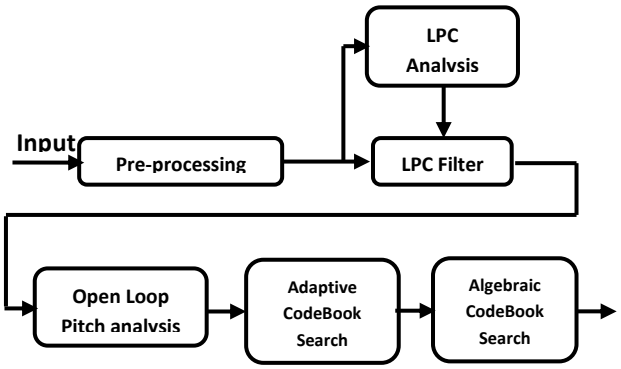

Fig. 4: the Basic block diagram of the AMR Coder [16]

The first stage is the pre-processing stage it consists of a high pass filter and a down scaling to reduce the system complexity. Second stage is the 10th order linear prediction analysis it is done twice per frame for the full rate at 12.2 Kbps. Next parameter to be extracted in this stage is the pitch of the sound; this is done using the open loop pitch estimation. The open loop pitch estimation is performed each half frame. The final two stages consist of the adaptive and algebraic codebooks. For AMR rates lower than 12.2 Kbps algorithms will be the same except some minor computations. For example for the AMR 5.15 Kbps the linear prediction analysis will be done once per frame not twice.

2.2. The PESQ (Perceptual Evaluation of Speech Quality) Algorithm

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 (5). It is standardized as ITU-T recommendation P.862 [17], [18], [19].

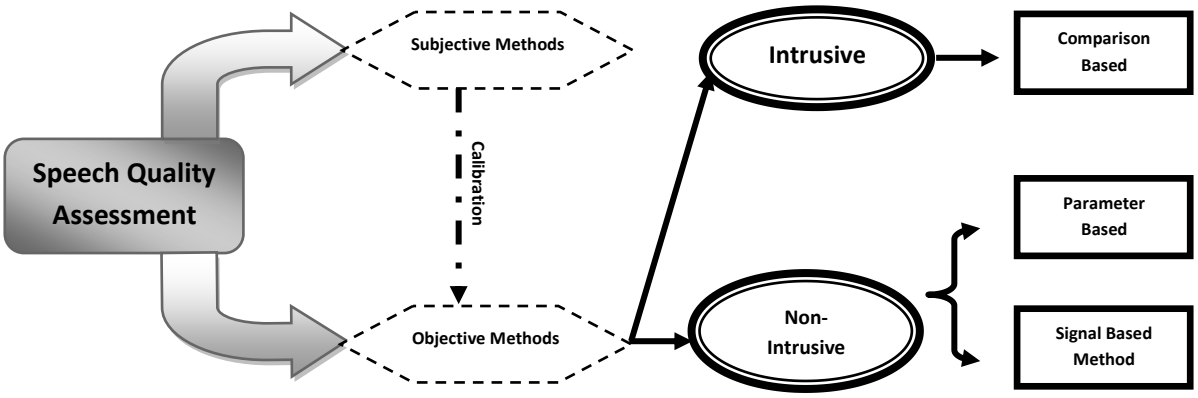

Fig. 5: Classification of speech quality assessment methods

PESQ compares X(t) [input signal] with Y(t) [output signal] which is passed through a communication system and evaluates the equivalent MOS (Mean Opinion Score) as in figure (6).

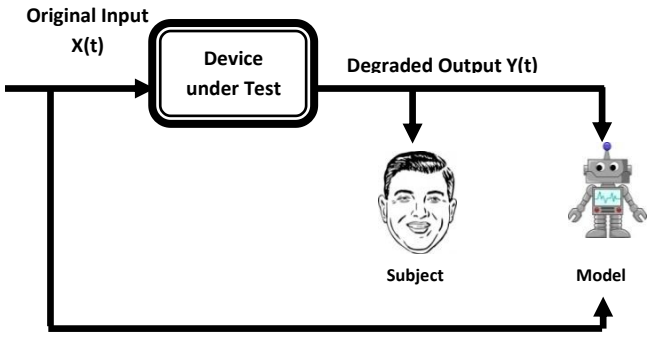

Fig. 6: PESQ basic philosophy.

The first step of the PESQ algorithm is to use a series of delays between the original signal and the degraded one to get the actual time delay between both. This step is called the time alignment algorithm. The algorithm can handle delay changes during both silence and active speech. Based on the set of delays that are found the PESQ compares the original signal with the delayed one using a computer model as illustrated in the figure (6).

The computer model replaces the subject (human). This model actually consists of two models. The first is the perceptual model that responsible for extracting the speech parameters and the second is the cognitive one that makes the actual judgement. These models are used to compare the input and the output of the device under test.

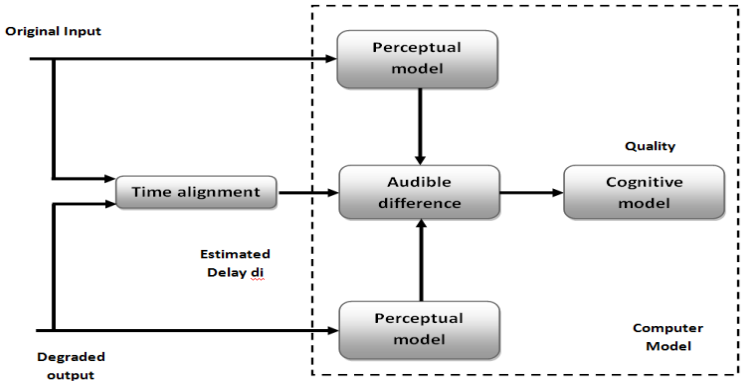

Fig. 7: the Basic block diagram of the PESQ algorithm

2.3. Results Analysis

The test strategy can be summarized as follows: First 10 voice files recorded [from the same person] for each of the English, Arabic and Cairo accent. This process repeated many times with different persons where the duration of each audio file is 10 seconds. Sec these files used to generate the output of the AMR Full rate 12.2 Kbps using standard Ericson tool, ACELP G.723.1 implemented using MATLAB and G.711 SIMULINK model. Third, the using of the PESQ algorithm (using Matlab code) to estimate the quality, finally tabulate the results and conclude the effects.

2.3.1. The Effect of the Spoken Language on the Waveform Coders and LP Coders.

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 (waveform) and G.723.1 (LP) coder and decoder. The decoded speech undergoes the PESQ test.

The following figure shows the relation between the SNR (Signal to Noise Ratio) for 10 speech samples (10 files [with different languages – English, Arabic and Cairo accent] each file 10 seconds duration) for both G.711 and G.723.1. The SNR is given by:

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

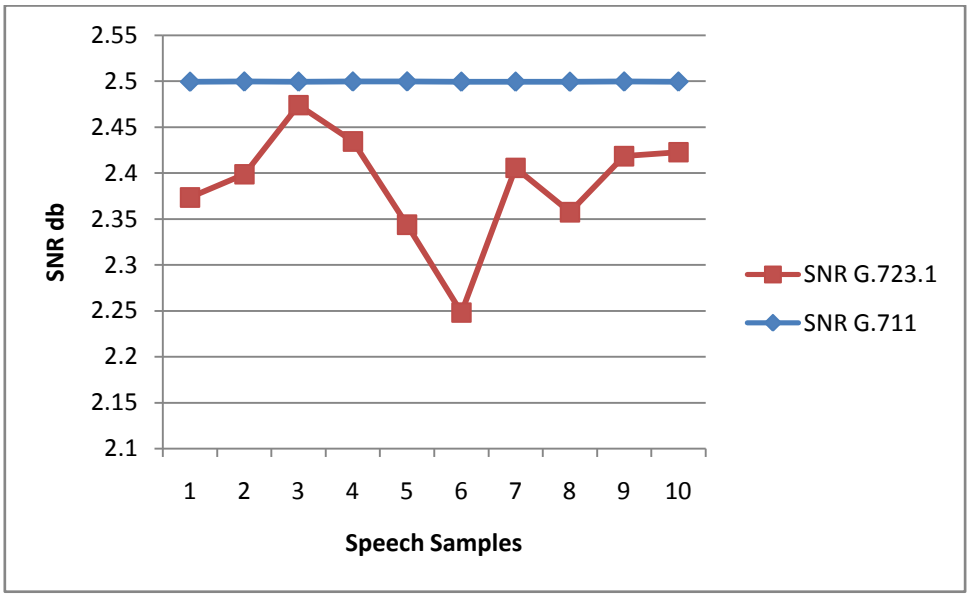

Fig. 8: SNR for G.711 and G.723.1 Vs sample number

It can be clearly concluded from the figure that the G.711 (waveform coder) does not affected by the change in the spoken language. This is actually because the waveform coders depend only on the samples amplitude in encoding and reconstruction of the speech while the LP coders depend on the extraction and reproduction of the signal parameters.

Table (1) and figure (9) illustrate samples of the PESQ scores for G.723.1 (Arabic, English and Cairo accent). For better estimation of the language or accent change effects the standard deviation has been used. 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 (2) shows the standard deviation for two males and two females PESQ scores

.

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

Table 1: PESQ MOS Score for G.723.1

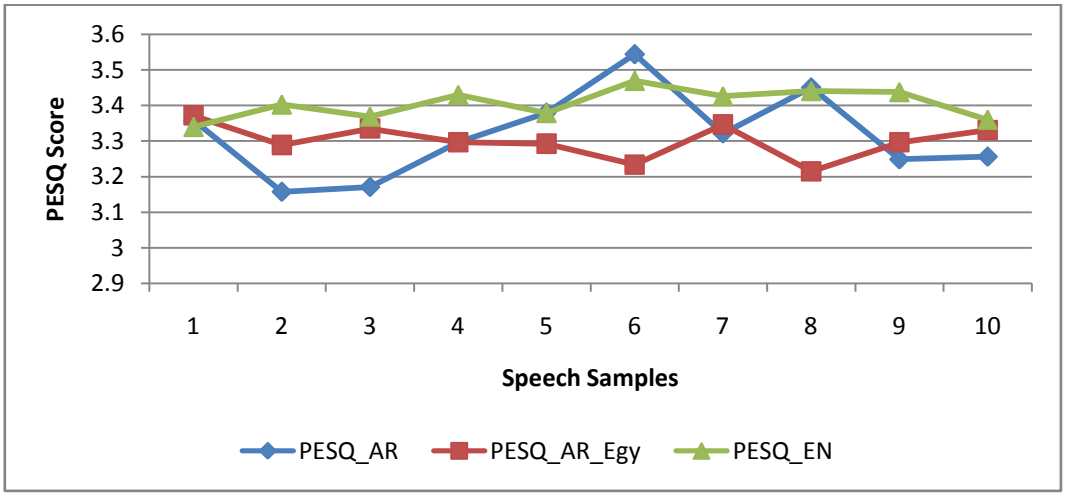

Fig. 9: PESQ MOS Score for G.723.1 versus Sample number.

	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 2: The Standard deviation values for the sample PESQ output

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

1. The PESQ score for English recorded speech is very slightly higher for most cases than either Arabic or Cairo accented Arabic.
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 the figure.
3. The above result is also confirmed by the lower standard deviation for English than either Arabic or Cairo accented Arabic, table 2.
4. The G.711 is always better than G.723.1; however this will be in the expense of the band width and the total cost.

Finally it has been proved that the waveform coders did not affected by the change in the spoken language while the LP coders did.

2.3.2. The Effect of the Spoken Language and the increase of the Compression ratio in LP Coders.

In the last section it has been proved that the quality of the speech affected when the spoken language is other than English for the LP coders. In this section we will prove that this quality will be affected more with the increase in the compression ratio of the coder. Table (3) and figure (10) indicates the comparison between G.723.1 5.3 Kbps (English and Arabic) and AMR Full rate 12.2 Kbps for English. This comparison is based on the PESQ values.

G.723.1 PESQ for Cairo accent	AMR PESQ for EN	G.723.1 PESQ for EN
3.3715	3.7799	3.395
3.2888	3.8234	3.4025
3.335	3.7834	3.4209
3.2966	3.7567	3.396
3.2924	3.7626	3.3787
3.2338	3.7903	3.4194
3.346	3.7533	3.4259
3.215	3.724	3.4407
3.2963	3.7276	3.4376
3.3306	3.7732	3.3897

Table 3: PESQ Values for Cairo accent, English with G.723.1 and English with AMR CODECs

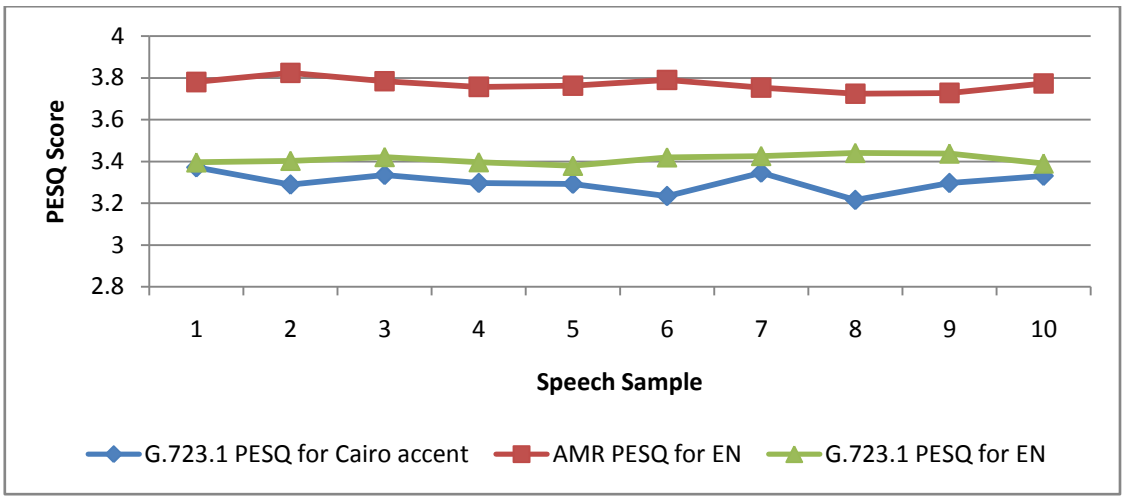

Fig 10: PESQ Curves for Cairo accent, English with G.723.1 and English with AMR CODECs

It can be shown that, the Cairo accent (coded with G.723.1) is lower in quality and higher in instability from both AMR with English and even G.723.1 with English of the same bit rate. Figure (11) shows the standard deviation for the PESQ results of different languages, which confirms the observation of the more instability with Arabic and Cairo accent.

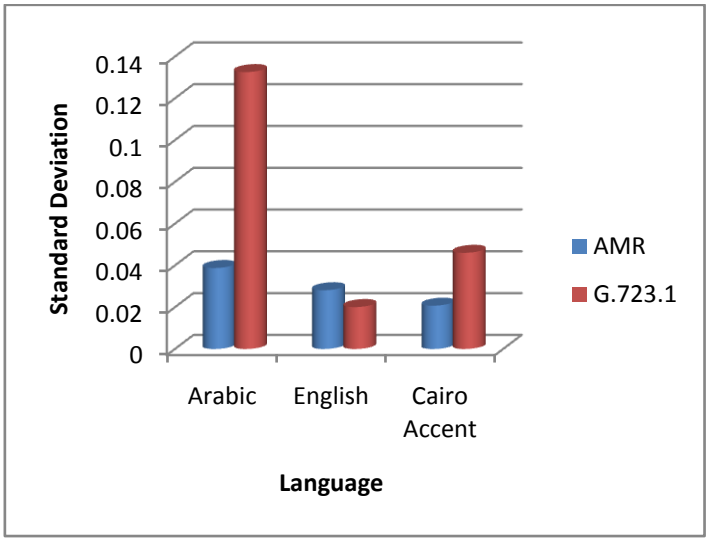

Fig 11: The Standard deviation values for the PESQ outputs for different languages.

3. The Problem Identification.

The efforts discussed in the last section were to prove the problem existence, it is now the time to identify the problem or in another word to find the defective part of the coder. To demonstrate the real problem a test on the speech files recorded and the LP vector quantization codebook has to be done. This section organized as follows, the first section describes briefly the basic concepts of the LBG (Linde Buzo Gray) algorithm and the second section gives the basic idea of the MFCC (Mel Frequency Cepstrum Coefficients) while the third section introduces the practical work test and results.

3.1. Linde Buzo Gray Algorithm (LBG).

LP coefficients must be quantized before transmission. The LP quantization is done using vector quantization which depends on vectors tabulated in the codebook used. Many algorithms used for codebook generation like:

- LBG algorithm.
- K-mean algorithm.

LBG is used here for the investigation of the problem of "The effect of the change in the spoken language on the parametric coders". The following flow chart describes the LBG algorithm [19].

As the flow chart indicates we start by defining the size of the CodeBook N or NxM if it was multidimensional CodeBook. Then select a random data from the input data source as the initial CodeBook, the next step is to measure the Euclidean Distance (Eq 1) measure clusterize the vector around each codeword.

$$d(p,q) = \sqrt{(p1-q1)^2 + (p2-q2)^2 \ \ldots + (pn-qn)^2} \qquad \text{Eq 3}$$

The final step is to compute the new set of codewords by obtaining the average of each cluster:

$$Y(i) = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \qquad \text{Eq 4}$$

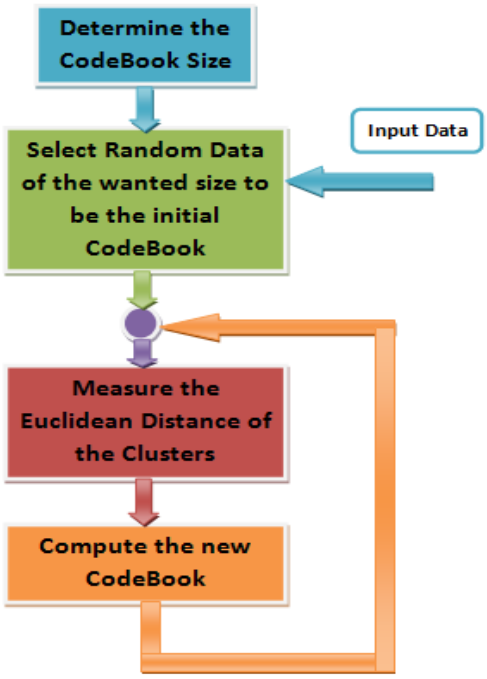

Fig 12: LBG algorithm.

Where:

- i is the component of each vector.
- m is the number of vectors in a cluster.

The last two steps will be repeated until no change in the codewords.

3.2. Mel Frequency Cepstral Coefficients (MFCC)

The Mel frequency scale is given by [20]:

$$F_{mel} = \frac{1000}{\log(2)}\left[1 + \frac{F_{Hz}}{1000}\right] \qquad \text{Eq 5}$$

And Mel Frequency Cepstral Coefficients (MFCC) derived using filter bank as in the figure below [20]:

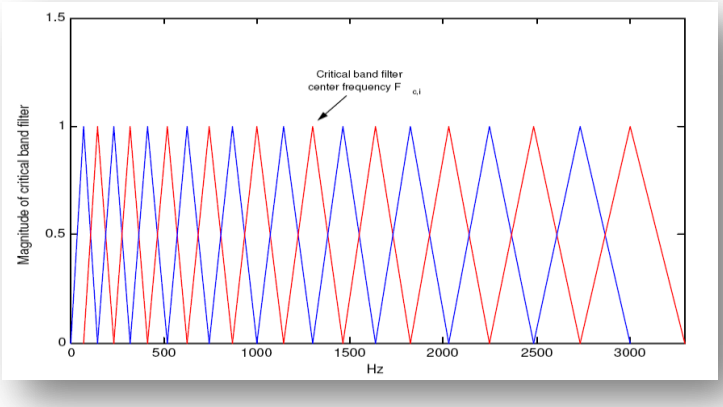

Fig 12: Mel scale filter bank.

It has been found that the energy in a critical band of a particular frequency influence the human auditory system perception. These bands linear below 1 KHz and logarithmic above 1 KHz. Combined together we can find the MFCC by:

$$MFCC(i) = \sum_{k=0}^{N/2} \log|s(n)| \, . H_i(k \frac{2\pi}{M}) \quad \text{Eq 6}$$

Where:

- N is the frame length.
- S(n) is the Discrete Fourier Transform (DFT) of the input signal.
- H_i is a critical band filter
- M is the number of points used in the DFT.

3.3. Results Analysis.

The practical work done here is very simple but of a great importance and can be summarized as follows:

1. Use MFCC equation to generate the MFCC coefficients for the recorded speech files of English, Arabic and Cairo accent (using Matlab code).
2. Draw the MFCC 5^{th} with 6^{th} coefficients for each language or accent.
3. Use the LBG algorithm explained before (Matlab code) to get the English CodeBook.
4. Draw a figure for the CodeBook with English, Arabic and Cairo accents MFCC and observe the quantization errors.
5. Measure the error using the distortion between the CodeBook and MFCC coefficients for English, Arabic and Cairo accent.

Figure 13 shows the output of the MFCC for each of English, Arabic and Cairo accent. As it can be seen from the figure there is a distribution difference between the English with Arabic and English with Cairo accent, although there is a similarity between Arabic and Cairo accent. This may be because the dedicated phonemes of Arabic and Cairo accent that does not exist in English language.

The second step as explained before is the use of LBG algorithm to generate the English CodeBook and compare it with MFCC from English, Arabic and Cairo accent as in figures from 14 to 16. It is clear from these figures that there are areas that have not any corresponding (very near) values in the CodeBook (circulated areas). This will affect the total distortion and hence the quality of the CodeBook and the overall coder.

The distortion is the average value of the difference between the exact MFCC and the CodeBook value used. Figure 17 shows that the CodeBook tends to have higher distortion with Arabic and Cairo accent over English which confirms the tests results in section 2.

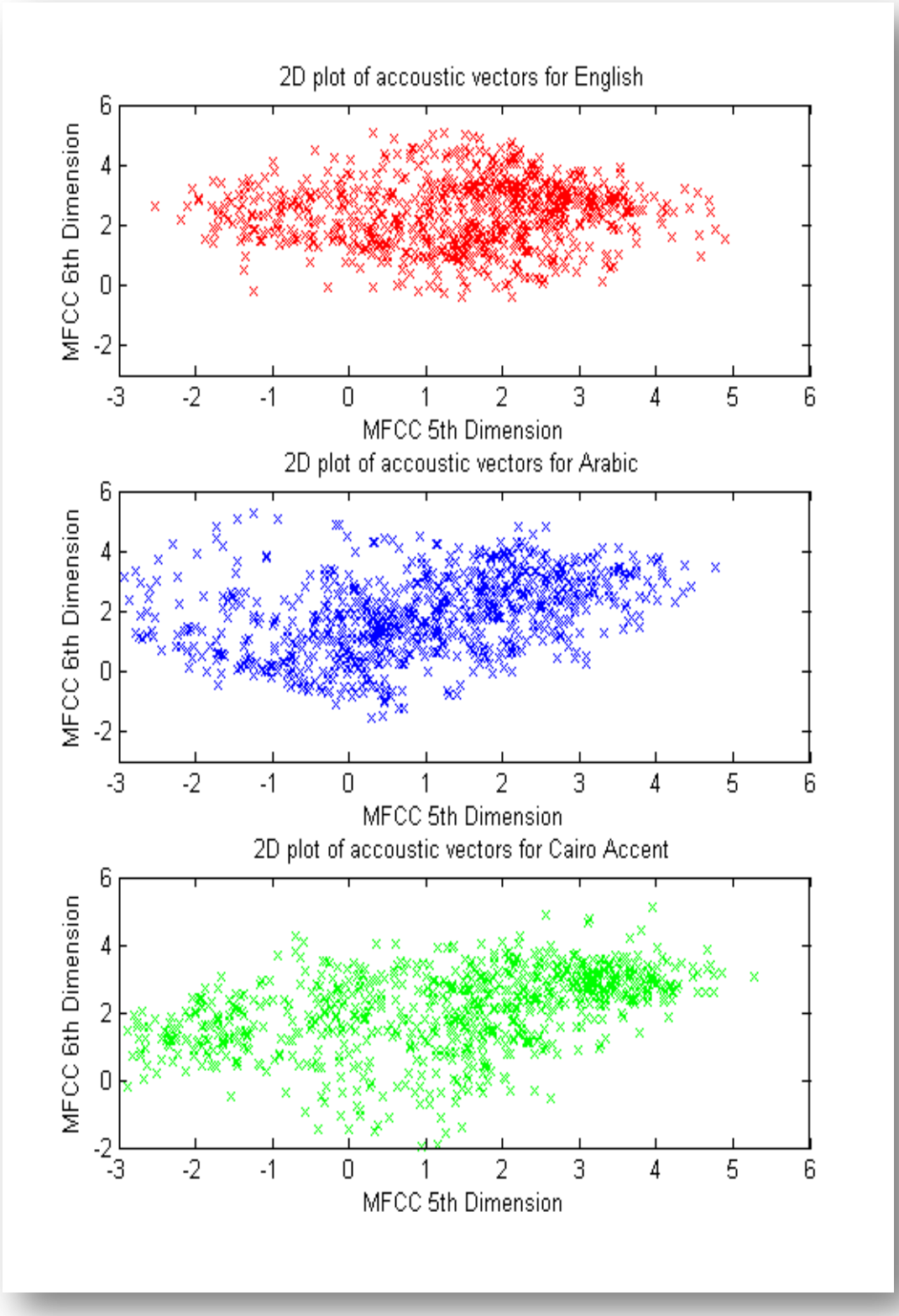

Fig 13: 2D plot of acoustic vectors for English, Arabic and Cairo accent

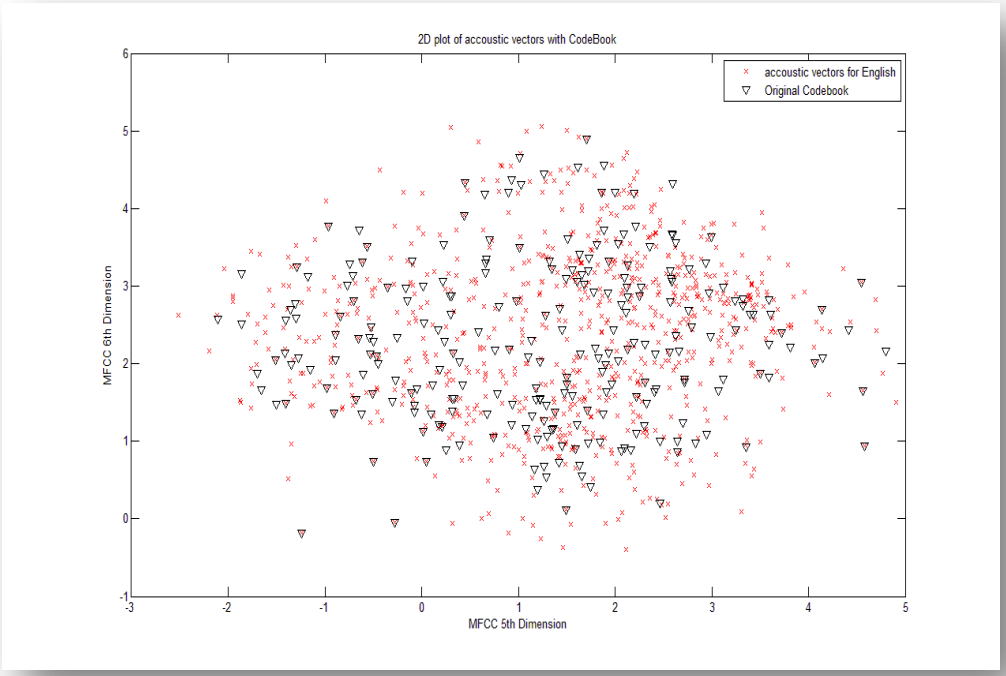

Fig 14: CodeBook with English MFCC

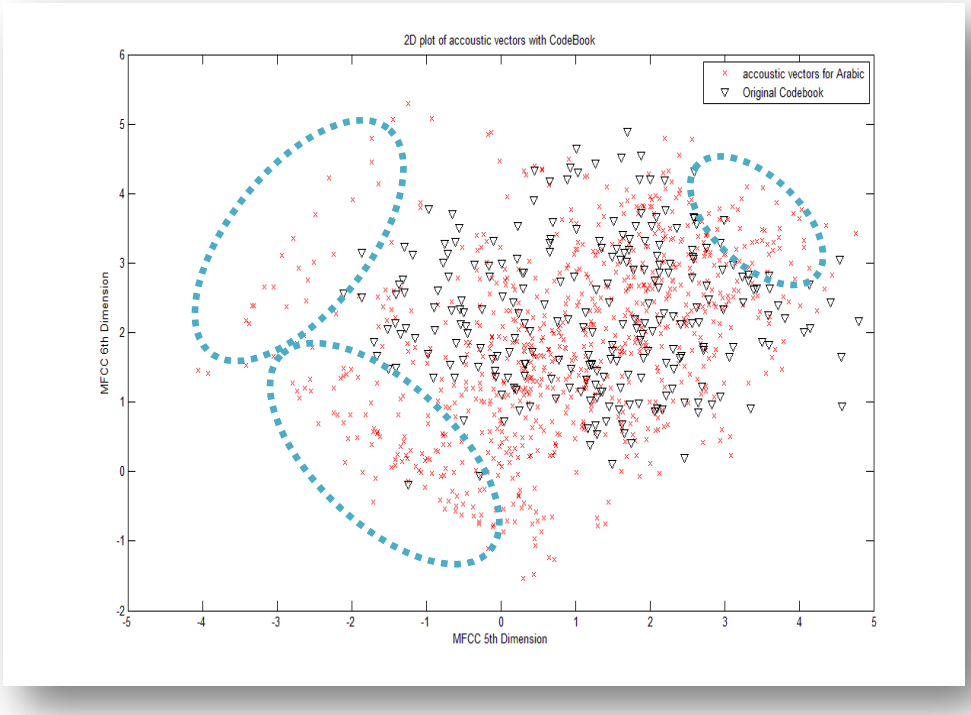

Fig 15: CodeBook with Arabic MFCC

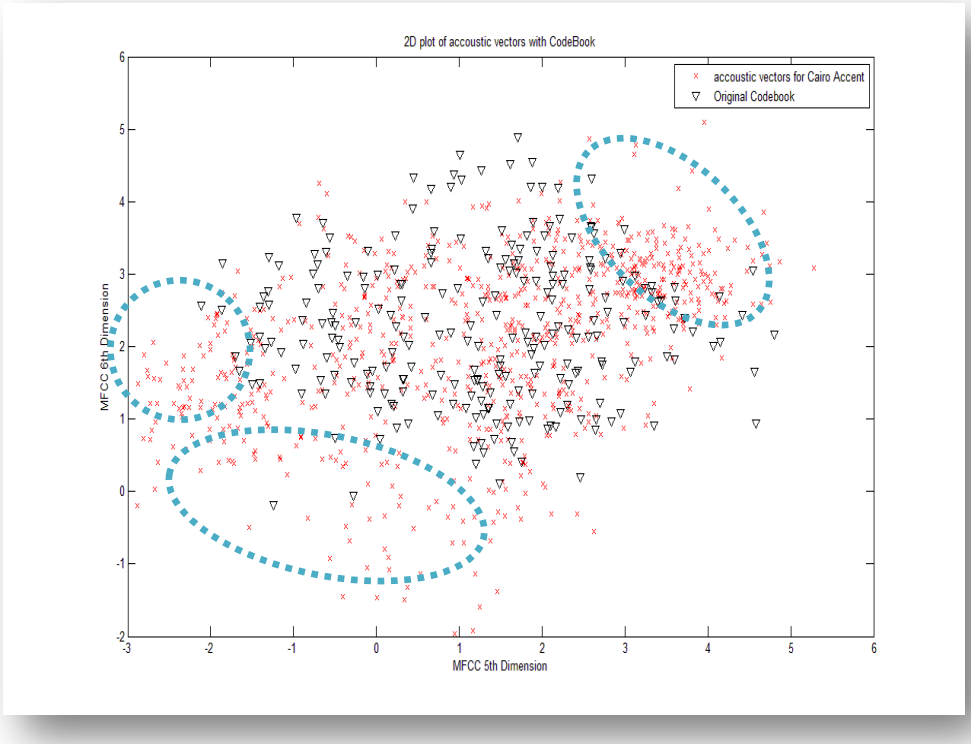

Fig 16: CodeBook with Cairo accent MFCC

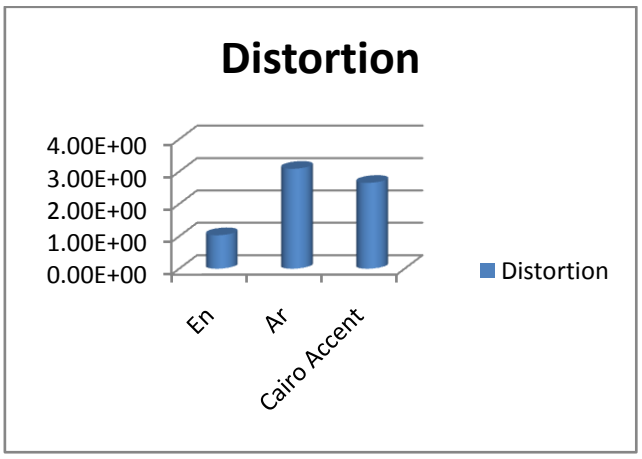

Fig 17: Distortion measure for English, Arabic and Cairo accent.

4. The Conclusion and Future Work.

The final conclusion of this paper is summarized in the following points:

1. The LP coders or parametric coders affected by the change in the spoken language or accents while the waveform coders did not. It is shown that these coders when worked with English language gives better PESQ score and higher stability than with Arabic or Cairo accent. This conclusion proven by the comparison between the G.723.1 and G.711 coders.

2. It has been proven that the degradation in the speech quality becomes greater when the compression ratio increased with the change in the spoken language or accent. This conclusion proven by the comparison between the G.723.1 and AMR coders.

3. The main problem in the parametric coders is in the LP quantization CodeBook. This has to be optimized according to the mother language for the coder to be used with.

A better quality will be obtained when using a new CodeBook generated for Arabic language or mixed codebook for many languages and this will be the future work.

REFERENCES

1- *Behrouz A. Forouzan*, "**Data Communication and Networks**", 2nd edition, by McGrew Hill, 2002.

2- *WAI C. CHU*," **Speech Coding Algorithms Foundation and Evolution of Standardized Coders**", by John Wiley & Sons, Inc, 2003.

3- *Jeremy Bradbury*, "**Linear Predictive Coding**", Mc G. Hill, 2000.

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, 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 Languages on Low Rate Speech Coders**", In Proceedings of Fourth International Conference on Spoken Language Processing, ICSLP 96, 1996.

7- *Mohamad Itani, Šarūnas Paulikas*, "**Influence of Languages on CELP CODECs Performance**", INFORMATION TECHNOLOGY AND CONTROL, 2008.

8- *Mohamad Itani, Šarūnas Paulikas*, "**Lithuanian Speech Records Database for Voice CODECs Quality Assessment**", INFORMATION TECHNOLOGY AND CONTROL, 2010.

9- *Paulikas, S.*" **Assessment of CELP CODECs Quality in Multi-lingual Environment**", IEEE Conference Publications, 2010.

10- *ITU – T Recommendation* **G.711** -1993.

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.

13- *Texas Instruments*, "**DSP C5000**", 2003 Texas Instruments.

14- *.Nokia Research Center*, "**Coding standards**", retrieved Mars 2010.

15- *3GPP*, 3GPP TS 26.073 " **AMR Speech Codec"**, Retrieved 2009-09-08.

16- *ITU – T Recommendation* **PESQ** P.862 (01) -1993.

17- *ITU – T Recommendation* **PESQ** P.862 (02) -1993.

18- *ITU – T Recommendation* **Quality of Speech Voice** P.830 -1997.

19- *Linde, Y., Buzo, A., Gray, R.M.*, "**An Algorithm for Vector Quantizer Design**", IEEE Transactions on Communications, vol. 28, 1980.

20- *Sahidullah, Md.; Saha and Goutam*. "**Design, Analysis and Experimental Evaluation of Block based Transformation in MFCC Computation for Speaker Recognition**", *Speech Communication* **54** (4), 2012.