TESPAR Coded Speech Quality Evaluation (TCSQE)

A New Algorithm for Objective Measurement of Speech Quality in Cellular Networks

Philip O. Adar
Solutions Consultant,
Telkom Kenya Limited.
P.O. Box 30301 – 00100,
Nairobi, Kenya.
padar@telkom.co.ke

Marcel O. Odhiambo
Department of Electrical and Electronic Engineering,
Faculty of Engineering,
The University of South Africa (UNISA),
P.O. Box Pretoria, South Africa.
Ohangmo@unisa.ac.za

***Abstract*—Speech transmission quality measurement in cellular networks is a major indicator of performance for end-to-end quality of service standards. Many approaches have been proposed in the previous studies, but the results correlation with subjective experiments still need further optimization, especially for quality determination using languages with unique phonetic structures i.e. clicking sounds. Moreover, the evaluation test data is always a key element in order to obtain representative and consistent results.**

In this paper, we introduce TESPAR coding technique for the design of a new algorithm for speech (voice) quality measurement in cellular/wireless telecommunication networks. Our experiments show that the results from this algorithm correlates well with subjective experiments using a variety of speech samples. The proposed algorithm is also computationally efficient than the existing methods and is suitable for quality measurement of longer speech signals than the current 8 seconds speech test data mostly in use today.

Keywords - speech quality; glottal excitations; TESPAR Codes; objective quality; subjective quality; Mean Opinion Score.

I. INTRODUCTION

Accurate computation of speech (or voice) transmission quality in telecommunication networks is done by either intrusive [1] or non-intrusive [2] techniques. The scientific measurement methods designed for this purpose model part of human auditory perception and are therefore called perceptual measurement methods. They detect and assess audible artefacts by comparing the degraded version of a signal e.g. output of the codec or a degraded speech sample received at one end of a telecommunication system, with the original signal. Some of these objective speech quality estimators in common use today [1-5] are complex in structure, require heavy computation and provide inefficient model of the loss of individual talker attributes during telephonic communications or coding impairments [6].

Some of the attributes of the individual speakers that differentiate one person's voice from another include each person's tone, emotional variations captured in speech, age and sex. These attributes are unique to individual speakers. The loss of these attributes would drastically erode the way in which people understand and relate to each other during telephonic communication. Think of having received a call over your cell phone, but the voice clarity is too poor that you have to keep on repeating yourself several times to be understood. Such kind of problems erodes the enthusiasm in communication.

It is therefore important that the telecommunication service operators maintain a high level of speech transmission quality to service subscribers. There is therefore a need to develop computationally efficient but accurate speech quality measurement algorithms that can be used to measure instantaneous status of the network quality in real time [6], [7], [8] and [9]. The aim is to allow optimization engineers an opportunity to consistently monitor and adjust network conditions for guaranteed Quality of Service (QoS) to end users.

In this study, a speech quality assessment algorithm which takes a pre-recorded original speech sample and compares this with a degraded version is developed. This degraded sample can be obtained directly at the receiving end of a communication channel, or may be a recorded output of a codec output. The feasibility and applicability of TESPAR coding [10] and [11] is explored to build a new algorithm which is accurate to different language attributes but also simple in structure, computationally efficient and cheap to deploy. Implementation and field testing of the developed algorithm is performed according to the procedures previously developed in [16].

The rest of this paper is organized as follows: section II discussed the speech quality evaluation process, section III discusses the TESPAR coding and features evaluation process. In section IV, the design and development process for TCSQE is given. Section V outlines the achieved performance measurement results and lastly the concluding remarks and recommendations are given in section VI.

II. SPEECH QUALITY EVALUATION

In [6], it has been shown that the most effective approach to quality of speech estimation is by the use of subjective techniques. The main reason for automated approaches which can be used repeatedly is due to large costs and time requirements with the use of human subjects. For more acceptable results, the developed scientific measurement approaches must also be accurate and consistent with the

benchmarked results from human judgments; [12] and [13]. Such metrics must also be economical (in both time and money) and be reflective of the speech intelligibility as perceived by human listeners. In [14] and [15], the objectives for perceptual speech quality estimator to be realized in this study were outlined. An engineering approach is proposed where the tasks to be accomplished are summarised as:

i. The design and development of a method for speech quality assessment, evaluation and prediction.
ii. The application of the developed method for automated speech quality assessment.
iii. The enhancement of quality evaluation systems under the aspect of speech/voice quality.

These three aims require the understanding of the processes of human perception and judgment with regard to voice and speech quality which is extensively discussed in [6].

III. SPEECH CODING AND FEATURES EVALUATION

The algorithm for TESPAR coding is implemented in two major phases: signal pre-processing and signal coding. Speech signals are first analyzed for end point detection. The phase comprising the beginning and end of active speech is detected and extracted for further processing [6]. After this, band pass filtering is done to remove frequency bands out of useful voice band frequency components that may be available in the signal. Since the fundamental frequency of a male's voiced speech ranges from 85 Hz to 155 Hz, and the need to design for wideband coding (WB-AMR as in 3G cellular Networks [16]), a band-pass filter of 70 Hz to 11 KHz is used. It has been tested and found that the TESPAR alphabet performs best at 11 KHz; [17] and [18]. The last procedure in signal pre-processing is noise suppression. This is done to remove the mean value of the signal from each of the sampled points to eliminate the DC offset. This is to allow the best comparison between samples [19].

The key approach in this coding algorithm is to pick-out the specific speech signal representations and transform this into a set of parameters for efficient analysis using statistical methods.

The analysis process starts by identifying the glottal excitation pattern and extract the relevant features of the speech signal. The potential problem with TESPAR coding is the representation of unvoiced speech. Suppose that $S_1(t)$ is the representation of the original speech signal and $S_2(t)$ is the representation of the degraded speech signal, then the distortion inherent in the degraded signal will be represented by $e(t) = S_2(t) - S_1(t)$. Where e(t) is the error signal due to noise and distortions along the communication channel.

TESPAR Coding algorithm gives A-Matrices and S-Matrices as the outputs. The A-Matrix is a p-by-p Matrix which takes the form of equation (1) while the S-Matrix is a p-by-1 Matrix whose general form is given in equation (2).

$$A = \begin{bmatrix} a_{(m,n)}, a_{(m+1,n)}, a_{(m+2,n)}, ..., a_{(p,p)} \\ a_{(m,n+1)}, a_{(m+1,n+1)}, a_{(m+2,n+1)}, ..., a_{(p,p)} \\ a_{(m,n+2)}, a_{(m+1,n+2)}, a_{(m+2,n+2)}, ..., a_{(p,p)} \\ . \\ . \\ . \\ a_{(m,p)}, a_{(m+1,p),} a_{(m+2,p)}, ..., a_{(p,p)} \end{bmatrix} \quad (1)$$

and

$$S = \left(s_{(m,n)}, s_{(m+1,n+1)}, s_{(m+2,n+2)}, ..., s_{(p,p)} \right) \quad (2)$$

where m = (1,2,3,…, p) and n = (1,2,3,…, p) and p is the size of the TESPAR alphabet codebook chosen. The acoustic correlates of speaker individuality are derived from the vocal-tract formants and glottal characteristics. The glottal excitation is itself the main source of excitation due to vocal-cord vibrations, therefore gives the best features of human voice/speech characteristics.

A. *Evaluating the A-Matrices*

The plot of the A-Matrices is visualised as an image mapping of the signal. A method commonly used to map features from a source corpus (i.e. original signal A-Matrix) to the features of a target corpus (i.e. degraded signal A-matrix) is based on Gaussian Mixture Models (GMMs) [20]. In A-Matrices, the feature vectors of the signal are assumed to be independent. Let $X = \{X_{(i,j)} \mid i = 1,2,\ldots,M \text{ and } j = 1,2,\ldots,M\}$ and $Y = \{Y_{(i,j)} \mid i = 1,2,\ldots,M \text{ and } j = 1,2,\ldots,M\}$ be the A-Matrices of the original and the degraded speech signals respectively, and let y_m denote the m-th acoustic class out of a total M classes. In a Gaussian mixture density model, the probability density functions of the K observations vector $x_k \ (k = 1, 2, 3, ..., K)$ is given by:

$$p(X_k) = \sum_{m=1}^{M} P(Y_m)\, p(X_k \mid Y_m) = \sum_{m=1}^{M} c_m N(X_k, \mu_m, \Sigma_m) \quad (3)$$

where X_k is an M-dimensional feature matrix. The conditional probability density for each M class Y_m is assumed to be a Gaussian component density $N\left(X_k, \mu_m, \Sigma_m\right)$, and c_m 's the related normalized mixture weights $\left(\sum_{m=1}^{M} c_m = 1\right)$ as shown in the previous studies in [20] and [21]. Each M-variate Gaussian component density is of the form:

$$N(X_k, \mu_m, \Sigma_m) = \frac{1}{\sqrt{|2\pi\Sigma_m|}} \exp\left\{-\frac{1}{2}(x_k - \mu_m)^T \Sigma_m^{-1}(x_k - \mu_m)\right\} \quad (4)$$

Where μ_m is n M-dimensional mean vector for the m-th component, and Σ_m is its covariance matrix. Therefore, the m-th mixture component is characterized by its mean (the centre) and the spread around that centre, expressed by the covariance matrix.

If we let $h_m(x) = P(Y_m \mid X)$ be the conditional probability of X belonging to class Y_m, and using equation 1 and Bayes' rule, we find:

$$h_m(X) = P(Y_m \mid X) = \frac{P(Y_m)\, p(X \mid Y_m)}{p(x)} = \frac{c_m N(X, \mu_m, \Sigma_m)}{\sum_{i=1}^{M} c_i N(X, \mu_i, \Sigma_i)} \quad (5)$$

B. Evaluating the S-Matrices

Let $X = (x_1, ..., x_p)^t$ and $Y = (y_1, ..., y_p)^t$ each be a P-dimensional TESPAR S-matrix vector of the original and degraded signals respectively. Each of the components of X and Y are univariate random variables $X_j (j = 1, ..., p)$ and $Y_j (j = 1, ..., p)$ which should both be the same if the original signal and the degraded signal are similar (i.e. no distortion). The distribution functions of both X and Y components are also invariant with respect to translation of the time axis, as they are obtained from the coded sequence of an initially non-stationary speech signal.

The evaluation on the S-matrices is performed as a predictive approach to estimate the deviation between the S-Matrix of the processed signal from the S-Matrix of the original signal. From the foregoing, there are two sets of data "points" (i.e. X_1 and X_2), defined by the numerical attributes from the two observations.

The method to estimate the degradation in speech from the S-matrices is to map the features extracted from the "original/clean speech" sample with a set of features extracted from the "degraded/processed speech" sample. That is, the method is to find a mapping between the two sets of vectors received from codebook processing. This is adapted from [20]:

$$S_i^{(A \to B)} = \frac{\sum_{j=1}^{J} w_{i,j} S_j^{(B)}}{\sum_{j=1}^{J} w_{i,j}} \quad (6)$$

here A and B represent source and target sets respectively, $S_i^{(A \to B)}$ is the i-th spectral vector of the mapped source codebook, and $W_{i,j}$ are the histogram counts of how often vector i of codebook A was found to correspond to vector j in codebook B, itself containing j vectors S_j^B. Where MS is the mean similarity value between S_1 and S_2, $\bar{S}_1$ and $\bar{S}_2$ are the statistical means of S_1 and S_2 respectively.

IV. TCSQE DESIGN AND DEVELOPMENT

The TCSQE speech quality perception algorithm is implemented through a computational algorithm that allows efficient extraction of the glottal excitation characteristics which best represents vocal tract geometry.

TCSQE is designed to be an intrusive end-to-end objective speech quality measurement algorithm. The algorithm therefore, requires that, both the original and the degraded speech signals samples be available for quality evaluation. For accuracy in signal evaluation, it is required that both the original and the degraded signals be correlated in time. Thus, time-delay compensation algorithm [22] is used to align the two signals before quality computation. The general structure of TCSQE algorithm is given in figure 3 and has been implemented in [6].

In this study, the implementation has been done using MATLAB programming environment. The Mean Opinion Score (MOS) is calculated from equation (7) (which is developed in [7]). The test results are given in section V.

$$MOS = \frac{1+4}{1+\exp(-1.7244 * x + 5.0187)} \quad (7)$$

V. PERFORMANCE MEASUREMENTS

(a) Performance of objective estimators with consideration to accuracy

Figure 1 shows the plots of PESQ [1], MNB [3], TCSQE [6] and EMBSD [23] with a given set of speech data samples. The graphs indicate the results of the various objective quality estimates in MOS scale associated with each test conducted. Because objective quality measures generate a distortion number, the MOS scores plotted are already transformed by the functions described in [24] and is therefore directly comparable to subjective MOS results. The comparison of performance of each objective estimator is therefore direct as the output results fall on the same scale (i.e. MOS [24]).

The objective evaluation of PESQ, MNB, EMBSD and TCSQE were conducted again, but this time round, with longer speech samples. Figure 2 depicts the results obtained from the evaluations. The performance metric of the objective quality measures against subjective ratings for both MOS scores and the MOS difference, were calculated using the transformed objective estimates. These results are summarized in table I [6].

Table I: Correlation coefficients and Standard Error of Estimates (SEE) of objective quality measures with speech from shorter sentences (i.e. 8 seconds and below)

Objective Estimator	Correlation Coefficient		Standard Error	
	MOS	DMOS	MOS	DMOS
PESQ	0.98	0.98	0.22	0.21
MNB	0.91	0.92	0.14	0.14
EMBSD	0.89	0.91	0.13	0.13
TCSQE	0.97	0.98	0.15	0.15

According to the results of table I, the best correlation coefficient is achieved with PESQ for both MOS and DMOS experiments. Other good performances were realized with TCSQE, MNB and EMBSD, in that order. The correlation coefficient and SEE calculated again for objective and subjective results with a different data set of speech samples as shown in table II.

Table II: Correlation coefficients and SEE of objective quality measures with speech from longer sentences (i.e. 30 seconds or more).

Objective Estimator	Correlation Coefficient		Standard Error	
	MOS	DMOS	MOS	DMOS
PESQ	0.97	0.96	0.22	0.21
MNB	0.92	0.94	0.14	0.14
EMBSD	0.92	0.96	0.13	0.13
TCSQE	0.98	0.98	0.15	0.15

The performance of TCSQE clearly shows an improvement over PESQ, MNB, EMBSD over test with long speech samples.

Longer speech samples were one of the most challenging data sets for objective quality measures. For practical cellular telephony usage, the channel conditions are not static. The conditions keep varying depending on instantaneous network load, signal strengths, handover and so on. The use of short speech data lasting 10 seconds or less can therefore not provide accurate average user perception since this is just a very short time frame for a meaningful conversation. Normally, many tests have been previously conducted by short speech samples but for accurate evaluation of transmission channels, longer speech data would present accurate user experiences while using live telecommunication systems.

Though the test results of both short and long speech data sets are still low, the performance of the objective quality measures for longer speech data can be considered as the most accurate of these measures when they are actually applied to the real network applications. Although the performance of TCSQE was still not satisfactory for real network application with short speech data sets, it certainly shows promising results for longer speech samples.

(b) Performance of objective estimators with consideration to computational complexity

Apart from the length of data signals and accuracy of objective estimators, computational efficiency is also an important factor to consider when choosing among many objective estimators for a particular purpose. This is necessary as it determines the total time required for conducting tests and evaluation on telecommunication networks. Since objective estimators generally approximate human judgmental processes, the accuracy is improved with a large number of testing. If each testing session requires a substantial amount of processing power and time, this could impact on the cost of required quality procedures and expenses for commercial telecommunication operators.

For this reason, many operators would therefore prefer objective systems which are efficient in both cost and computation while maintaining the desired accuracy. Table III gives the computation time for each of the objective estimators used for both long and short speech data samples.

Table III: Average computation time in seconds for Data Sets I and II

Objective Estimator	Short speech samples (<8sec)	Short speech samples (>30sec)
PESQ	22	58
MNB	20	55
EMBSD	15	48
TCSQE	16	50

The evaluation times recorded in table III is a composition of both the data transmission and acquisition processes during live network measurement. Apart from the time due to signal propagation (which is constant for all cases), the total time taken to compute the final result is dependent on the complexity of the algorithm and functions that have to be computed. Therefore, it can be noted from table III that TCSQE processing time is much shorter when compared to the other objective estimators. This can be attributed to the evaluation process of TCSQE. Instead of using all samples under test, the generation of TESPAR codes significantly reduces the number of sample points per speech sample while maintaining the desired quality. It has been established in [17] and [18] that the TESPAR coding process retains enough information about speech signals which may also be used for speech recognition purposes with considerable accuracy as applied elsewhere in speech recognition applications.

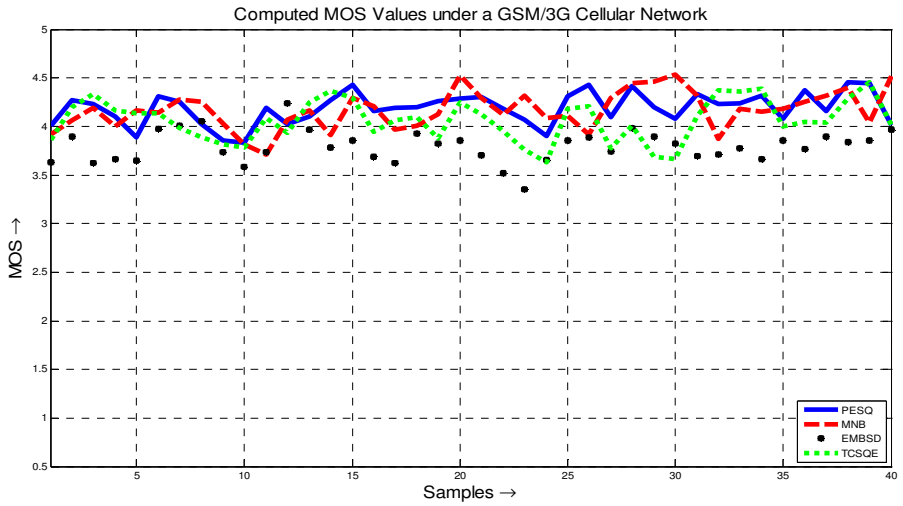

Figure 1: The graphs showing results of different objective speech quality estimators in comparison to TCSQE with short speech samples (i.e. 8 seconds long).

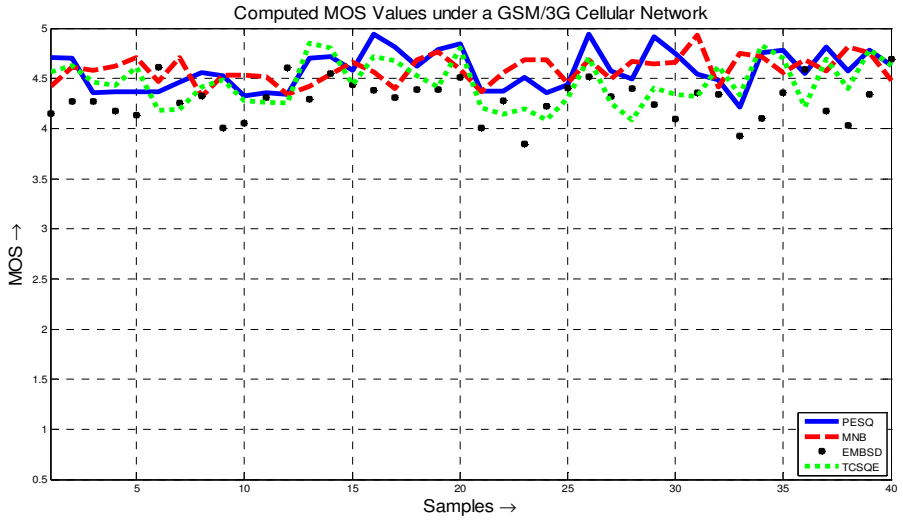

Figure 2: The graphs showing results of different objective speech quality estimators in comparison to TCSQE using longer speech samples as test data (i.e.30 seconds long).

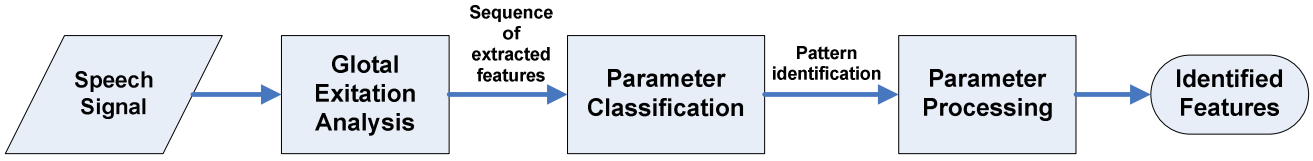

Figure 3: Speech signal features verification process

VI. CONCLUSIONS AND RECOMMENDATIONS

The performance of TCSQE with different speech data samples were evaluated in this study. The results obtained show that the TCSQE performance is consistent after many independent test experiments (i.e. approximately 40 tests with each data set). Such consistency is a requirement for the reliability of objective speech estimators. This puts the TCSQE at the same level with the existing standard techniques.

Accuracy is also of paramount consideration of perceptual quality evaluation methods. The best way to establish the accuracy of an algorithm is to calculate the correlation between the subjective tests experiments with objective results obtained after many independent tests. It has been ascertained (table I and II) that TCSQE achieves a correlation measure of up to 0.98 with test data (a) and 0.92 with test data (b). It is with the evaluation of a longer speech signal (i.e. data (b)) that TCSEQ shows a significant improvement over PESQ.

TCSQE has a minimal dependency on the perceptual transformation of speech signals using Fourier Transform. Instead, a straight forward coding procedure is used for discriminant analysis of speech samples under evaluation. This has enabled TCSQE to have a reduced computational complexity than any of the existing algorithms.

The method of evaluating voice features after coding is based on GMM of TESPAR A-Matrices. Though the TESPAR coding process delivers accurate speech feature extraction process, further work using machine learning algorithms can be exploited for feature analysis and MOS computation. This may further lower the computational complexity and improve the algorithms accuracy.

REFERENCES

[1] ITU-T Recommendation P.862. 2001. *Perceptual evaluation of speech quality(PESQ): An objective method for end-to-end speech quality assessment of a narrow- band telephone network and speech codecs*. International Telecommunications Union - Telecommunications Standardization Section. Geneva: Switzerland.

[2] ITU-T Recommendation P.563. 2004. *Perceptual non-intrusive single-sided speech quality measure*. International Telecommunications Union, Geneva, Switzerland..

[3] Stephene Voran. 1999. Objective estimation of perceived speech quality – Part I and Part II: Development and Evaluation of the measuring normalizing block technique. In: *IEEE Trans. on Speech and Audio Processing*, vol. 7, July: 371–390.

[4] Berger, J. 1997. TOSQA – Telecommunication objective speech quality assessment. *ITU-T SG12 COM-34E, Dec*. 1997.

[5] Holub, J. 2007. ETSI Workshop on Speech and Noise in Wideband Communication European Telecommunications Standards Institute 22nd-23rd May, Sophia Antipoli France.

[6] Adar, P.O. 2008. Optimal Measurement of Speech Transmission Quality in GSM/3G Cellular Networks. Master thesis, Tshwane University of Technology, June 2008.

[7] John, C.M. & Irina, C.C. 2004. Mapping of objective voice quality metric to a MOS domain for field measurements.

[8] Rix, A. W. 2003. Comparison between subjective listening quality and P.862 PESQ score. In: p*roceedings of online workshop measurement of speech and audio quality in networks*. Czech Republic, May: 17–25.

[9] EURESCOM STAFF Project P603. 1997. Quality of Service: Measurement method selection. *Technical Report, Deliverable 2*.

[10] Licklindder, J.C.R & Pollack, I. 1948. Effects of differentiation, integration and infinite peak clipping upon the intelligibility of speech. *Journal of the acoustical society of America*. Vol.20(1), Jan.:42-51.

[11] King, R. A. 2004. Waveform Coding Method. *United States Patent No: US6,748,354 B1*, June 8.

[12] ITU-T Recommendation BS.562-3. 1994. *Subjective assessment of sound quality*. International Telecommunications Union. Geneva: Switzerland.

[13] ITU-T Recommendation P.800. 1996. *Methods for subjective determination of transmission quality*. International Telecommunications Union, Geneva, Switzerland.

[14] Michael, C., Stefan, L., Kevin, L. M. & Robert, T. 1999. Can speech recognizers measure the effectiveness of encoding algorithms for digital speech transmission. National Institute of Standards and Technology, January 1999.

[15] Adar, P.O., Chatelain, D., Oyedapo, J., Kurien, A.M. 2007. Perceptual Speech Quality Measurement on Cellular Networks – A Proposal for Enhancing Objective Estimators. *In the proceedings of the 8th IEEE Africons International Conference(Poster Session)*. 26th -28th Sept.2007 Windhoek, Namibia.

[16] Adar, P.O., Chatelain, D., Oyedapo, J., Kurien, A.M. 2007. Optimal technique for speech quality evaluation on W-CDMA 3G cellular Networks". *In Proceedings of the 2007 IEEE 14th Int'l Conference on Telecommunications and 8th Malaysia Int'l Conference on Communications, ICT-MICC 2007,* Penang, Malaysia, May 14-17.

[17] Ryo Hwang. 1999. Speaker identification system using TESPAR technique. In: *Proceedings of the International Conference on Signal Processing Applications and Technology (ICSPAT '99).*Vol. 18, pp 445-453, Great Britain.

[18] George, M.H. & King, R.A. 1995. A robust speaker verification biometric. In: *Proc. of IEEE annual Conference on Security Technology*, UK, 18-20th Oct.: 41-46.

[19] Rodwell, G.M. & King, R.A. 1995. TESPAR/FANN Architecture for low-power, low-cost condition monitoring application", School of Engineering and Applied Science", Royal Military college of science, United Kingdom.

[20] Juergen Schoeter. Voice modification for applications in speech synthesis. AT&T labs research.

[21] Douglas A. Raynolds, Thomas F. Quatieri and Robert B. Dunn. 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal processing 10, 19-41, Academic Press, 2000.

[22] Rix, A.W., Hekstra, A.P., Hollier, M.P. & Beerends, J.G. 2002a. Perceptual Evaluation of Speech Quality (PESQ), the New ITU Standard for end-to-end quality assessment: Part I-Time delay compensation. *Journal of .Audio Engineering Society*, 50(10), Oct.: 755-764.

[23] Wonho Yang. 1999. Enhanced modified bark spectral distortion (EMBS): an objective speech quality measure based on audible distortions and cognitive model. *PhD Dissertation, Temple University, USA.*

[24] ITU-T Recommendation P.800.1. 2003. *Mean Opinion score Terminology*. International Telecommunications Union, Geneva: Switzerland.