

Talker Dependencies Factor in Analyzing Listening Quality on Perceptual Speech Quality Measurements

*Philip O. Adar and **Marcel O. Odhiambo

*Department of Business Sales and Wholesale, Telkom Kenya Limited,
P. O. Box 30301 - 00100, Nairobi, Kenya.

Tel: +254 (0)20 250 0806, Fax: +254 (0)20 323 2100

**Department of Electrical and Mining Engineering, University of South Africa (UNISA),
Private Bag X6, Florida 1710, South Africa.

Tel: +27 11 471 3141/3706, Fax: +27 11 471 3054

E-mail: padar@telkom.co.ke, ohangmo@unisa.ac.za,

Abstract— *Talker dependency attributes of the voiced speech such as pitch and harmonic characteristics have considerable importance and influence on speech quality. The loss of these attributes, especially on a speech signal transmitted through a communication network may render the received speech difficult to comprehend. This is one aspect of speech quality degradation which has not been exhaustively researched in the context of perceptual speech quality estimation in cellular networks.*

This paper proposes an enhancement algorithm that is useful as a refinement technique of the existing perceptual speech quality measurement algorithms. Laboratory and field experiments conducted in this study indicate that considerable improvements on Mean Opinion Score (MOS) correlations can be achieved by incorporating these new parameters as factors that influence overall speech quality.

Keywords: speech quality, MOS, quality determination parameters.

I. INTRODUCTION

Subjective methods of speech quality evaluation are considered classical and the most accurate [1]. With this measurement method, panels of human listeners are requested to gauge a set of speech data under test and give an opinion on the level of perceived degradation. A large number of these subjects, (i.e. at least 40) would be required for every testing experiments after which an average of all the results is calculated and the result would give the Mean Opinion Score (MOS) [2]. This method is associated with large manpower costs and requires lots of time, therefore not suitable for repeated quality determination in live telecommunication systems or laboratory speech coder/decoder performance determination experiments.

Objective measurement methods which closely mimic the human judgmental criteria have been proposed for automated experimental processes of determining speech quality. These scientific speech quality measurement methods have dominated active interest within the research community for several years now. The most dominant

application areas include Quality of Service (QoS) measurement in telecommunications networks by service operators and regulation authorities on one hand; and speech coder/decoder research and development community on the other. Among the most common intrusive measurement methods developed in the recent past include PESQ [3], NMB [4] and PEAQ [5]. These intrusive speech measurement methods are also popular due to their ease of implementation and accuracy as compared to the non-intrusive objective measurement approaches (e.g., [6]).

The PESQ algorithm gives MOS results which are closely related to subjective listening tests and is the most commonly used test algorithm. Among the test factors and applications for which PESQ has not been validated are temporal and amplitude clipping of speech, talker dependencies, listening level, loudness loss and multiple simultaneous talkers [3].

Figure 1 depicts the high level approach used in the logical design of these existing test algorithms [7].

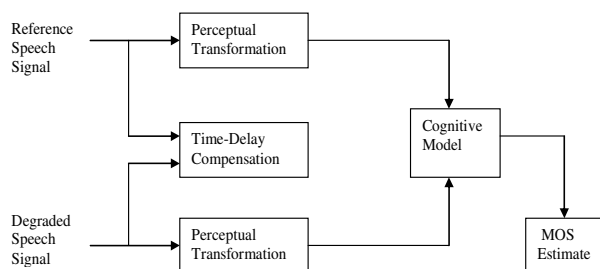


Figure 1: Block diagram for a speech quality estimator.

These objective quality prediction methods are implemented in machine executable algorithms which currently, cannot determine pitch variations, loss of individual voice cues and vocal/tone characteristics. These characteristics are in essence, very important tools in human languages and speech communication [7] and [8].

In [9], the existence of temporal discontinuities and their effects on delay estimates in evaluating quality of speech is investigated. It is observed that if temporal discontinuity impairment occurs in the midst of a syllable of speech after

initial delay compensation, it would very likely create some spectral distortion. Such errors when passed through to cognitive modeling would thus be evaluated unknowingly and this creates incorrect results.

Results achievable in such cases are inconsistent with human judgmental processes. Voice naturalness and pitch should be retained in a communication for such excellent quality score of 4.5 or more to exist [2]. This study strives to address the inconsistencies on objective intrusive algorithms through a study of harmonic analysis of the speech spectrum and pitch detection and tracking techniques. The additional information derived by the consideration of these speech attributes are implemented in a MOS Refinement Function (MRF) that is proposed in this study to improve on the dynamic accuracy of the objective speech measurement results.

The rest of this paper is organized as follows; section II discussed the speech modeling for quality determination algorithms and a selection of new parameters are studied. In Section III, an MRF is derived for implementation as an improvement to the existing algorithms. The tests and results are discussed in section IV; finally the conclusion and areas of further study are given in sections V and VI respectively.

II. SPEECH MODELLING AND PARAMETER CLUSTERING

The voiced signals of human natural speech have roughly the shape of a quasi-periodic impulse train. This means that the intervals between successive impulses are not exact, but vary slightly and the amplitudes of different impulses are not exactly constant. The mathematical description of a speech discrete-time, quasi-periodic impulse train, is given by:

$$x[n] = \sum_{m=-\infty}^{\infty} c_m \delta[n - n_m] = \begin{cases} c_m, & n = n_m \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The impulses occur at instant n_m and their amplitudes are c_m . The intervals are relatively sparse i.e., the differences $[n_m - n_{m-1}]$ are larger than 1. The reciprocal of the average interval between successive impulses is the *pitch*. The pitch is the characteristic of a person and defines the naturalness on someone's speech. Pitch of the speech signal, therefore forms the basis of voice classification in relation to age and gender of the speaker [10].

Voice quality is also defined and a composition of the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and super-laryngeal features and running continuously through the individual's speech. The distinctive tone of speech sounds produced by a particular person yields a particular voice, which must not be lost during communication.

The transformation of speech signals from the frequency domain to the time domain and vice-versa during transmission processes inherently induce some spectral losses to the speech signal. By selecting certain harmonics using some suitable criterion and comparing the losses versus gains with the degraded sample may present away to determine changes on the speech spectrum. Judgments of the quality of spectral estimates are based on estimation theory on the psychoacoustic models [11] of speech.

The following are the novel methods of speech classification that is relevant to track speech naturalness and pitch variations in objective quality evaluation algorithms.

(i) Coherence Function (CF), given by:

$$\Gamma^2(e^{j\theta}) = \frac{|S_{xy}(e^{j\theta})|^2}{S_x(e^{j\theta})S_y(e^{j\theta})} \quad (2)$$

where $|S_{xy}(e^{j\theta})|^2$ is the cross spectrum of two input signals $x[n]$ and $y[n]$, $S_x(e^{j\theta})$ and $S_y(e^{j\theta})$ are Energy Spectrum Density (ESD) of $x[n]$ and $y[n]$ respectively. Coherence function tends to 1 when $x[n]$ and $y[n]$ are clean speech signals. On the other hand, coherence function falls to 0 when $x[n]$ and $y[n]$ are uncorrelated noises.

(ii) Pitch Tracking on Voiced Speech (PTVS)

Speech analysis research uses pitch tracking to better understand how pitches are used in communication. Having a clear, reliable method of extracting these tonal patterns help one to understand how tonality is used in speech. While simple analysis is used for understanding certain aspects of speech (such as determining speech formants), pitch tracking provides a clear, more detailed picture of how pitch changes within speech. Pitch detection on successive wavelet approximations is used to determine and track the pitch of vocals [12].

The algorithm used in this work, based on time domain analysis, was designed to provide faster, more accurate method of pitch tracking than is possible with frequency domain methods. By employing the Fast Lifting Wavelets Transform (FLWT) [12] to simplify waveforms and then applying an intelligent peak-finding method, the period (and hence frequency) is accurately determined. Latency is reduced and response time improved.

The wavelet transform is similar to the Fourier transform in that it breaks a signal down into component parts. The Continuous Wavelet Transform (CWT) is based on the wavelet function, which is derived as:

$$\varphi_{s,\tau}(t) = \frac{1}{\sqrt{|s|}} \varphi\left(\frac{t-\tau}{s}\right) \quad (3)$$

Where φ is the mother wavelet, which serves as the basis of the waveform transform. Using equation (3) for a given signal $x(t)$, a set of wavelet coefficients $C_{s,\tau}$ is defined as;

$$C_{s,\tau} = \langle \varphi_{s,\tau}, X \rangle \quad (4)$$

$$= \sum_t \varphi_{s,\tau}(t) X(t) \quad (5)$$

$$= \frac{1}{\sqrt{|s|}} \sum_t \varphi\left(\frac{s-\tau}{s}\right) X(t) \quad (6)$$

These sets of coefficients provide the forward CWT. To get the original signal back, equation (3) and (6) are combined to derive the inverse CWT, which is given by:

$$x(t) = \sum_s \sum_{\tau} C_{s,\tau} \varphi(t) \quad (7)$$

Thus the CWT is arrived at, which decomposes a signal into a set of wavelet coefficients based on translations and dilations of the original mother wavelet.

(iii) Turbulent Noise Index (TNI)

The TNI is an acoustic measurement parameter which has been proposed to serve as a laryngeal function [13]. Experiments conducted with synthetic and natural voice show that TNI is almost independent from frequency and amplitude modulation noise. But, TNI is proved [14] to be dependent on the fundamental frequency and/or waveform amplitude variations in a voiced signal.

In calculating TNI, a voiced signal $v(i)$ is expressed as a sum of its quasi-periodic component $x(i)$, caused by vocal folds vibrations, and additive noise $a(i)$ (a periodic component), that represents the turbulence of air flow in the vocal tract , i.e.

$$V(i) = x(i) + a(i) \quad (8)$$

where $x(i)$ is 0 in the absence of vocal folds vibrations. For an interval of duration T_{\min} , equal to the minimal pitch period duration, the quasi-periodic component can be expressed as:

$$x(t_0+i+T) = kx(t_0+i), \quad 0 \leq i < T_{\min} \quad (9)$$

where t_0 is the beginning of a glottal cycle (transition from open to closed glottis), T is the pitch period duration at the moment t_0 , and k is a coefficient describing the change in the waveform amplitude of two consecutive periods. Both T and k are functions of the time variable t_0 and represent jitter (cycle-to-cycle frequency fluctuations) and shimmer (cycle-to-cycle waveform amplitude fluctuations) of the pitch period, respectively. So, the modulation noise has been taken into account in the term of $x(i)$ [8], and $a(i)$ represents only the turbulent noise. The additive component (turbulence noise) can be assumed to be Gaussian noise with zero mean value and stationary energy (during every period, its energy is approximately the same).

The derivation procedure for TNI is available in [11]. The formula gives TNI as a mean value for a voiced signal segment given by:

$$TNI = 100(1 - \bar{R}_{\max}) \quad (10)$$

where

$$\bar{R}_{\max} = \frac{1}{N-1} \sum_{n=1}^{N-1} R(t_n, T_n) \quad (11)$$

and $R(t_n, T_n)$ is the normalized autocorrelation function, calculated for the n -th glottal cycle beginning at t_n with a duration of T_n , and N is the number of the cycles.

In natural voiced signals the pitch period duration is relatively short (from 2.5 to 16.7 ms), so $R(t_n, T_n)$ oscillates around some mean value. When TNI is calculated as a mean value for a long enough segment of the signal, the result will not be influenced.

III. MODIFICATION OF THE CLASSICAL TESTING PROCEDURES

The derived voice quality indicators of section II are used to develop an MRF. This is at best subjected to the quality estimation algorithm after initial MOS estimates. Figure 2 illustrates a modification of the classical approach depicted previously in Figure 1.

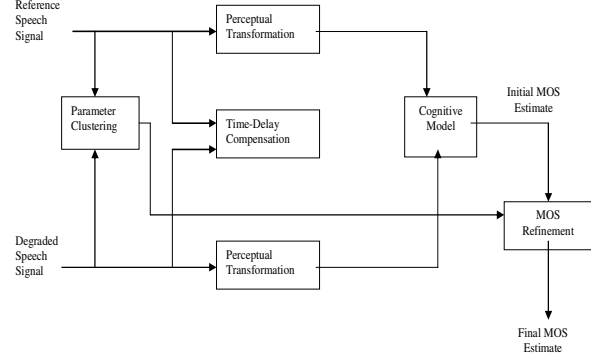


Figure 2: Proposed modification to speech

This MRF takes the form:

$$CPI = CF + PTSV - TNI \quad (12)$$

where CPI is the cluttered parameter index and CF, PTSV, TNI are the individual contributions of the Coherent Function (CF), the Pitch Tracking on Voiced Speech (PTSV) and the Turbulent Noise Index (TNI) respectively.

The MOS refinement is then achieved by using the following non-linear equation which is used to map CPI into the required MOS results. The equation is a modified relationship developed in [7] and is given by:

$$MOS_{refined} = 4.5 - 0.125 * CPI^{2.2} \quad (13)$$

where 4.5, is the expected maximum MOS achievable by objective estimators [1], the others (0.125, 2.2) are coefficients optimized for best performance. The overall effect of CPI on MOS is illustrated by the graph of Figure 3.

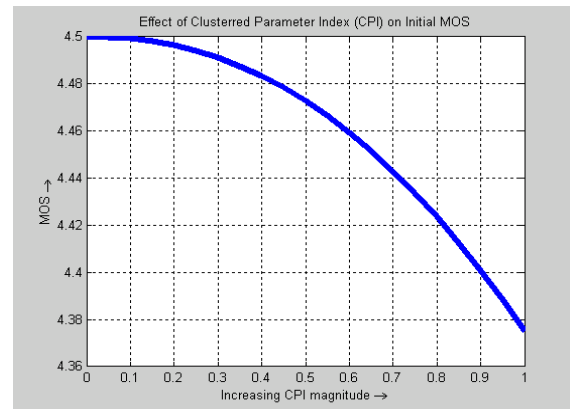


Figure 3: Effect of CPI on Initial MOS estimate

IV. EXPERIMENTS AND RESULTS

Speech quality ranking was achieved by comparing one or a combination of several metrics from the samples under tests. A dataset of 20 different short sentences of approximately 10 to 15 seconds long in the English language and a sample of other local African languages (10 male and 10 female) were used in this experiment. Original samples were transmitted through a cellular communication network. The experimental methods and results have been illustrated and discussed earlier in [7] and [8]. The received sample (degraded version of the original) was saved on a computer and evaluated using objective quality estimation algorithms (i.e. PESQ [3] and TCSQE [8]) implementation in Matlab.

On the received speech samples, speech quality measurement tests were run over them independently. The first was through the PESQ algorithm but with an extension to include the effect of clustered parameters (see figure 2). As expected, a decrease in the MOS values was noticed, in comparison to the initial test. The final testing involved an absolute category rating experiment using the degradation MOS [2]. 10 subjects (all students averagely in the age bracket of 23-30, but from different backgrounds) were selected. The listening instrument was the laptop windows sound card where the original, followed by the degraded version of the speech samples were played and the subjects asked to grade. The results were then averaged per sentences and plotted.

Figure 4 depicts the results generated from the different cases.

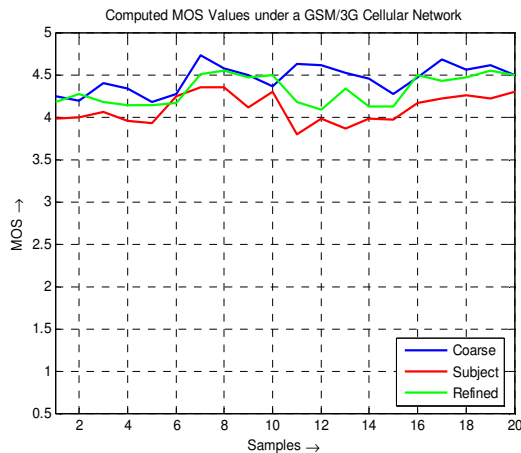


Figure 4: Computed MOS Values

Essentially, the differences in speech quality measures are depended on the inherent metrics between the two samples under test. The metrics used in quality computation relate to user perceptions which are obtained through subjective quality measurements [15]. Since speech communications quality is never a static entity, but happens in a specific situation for each user individually, objective results are benchmarked with subjective experiments before a reliable quality prediction measure can be given.

The correlation analysis on the results obtained in Figure 4 was performed. The Pearson's correlation method was used. It was observed that subjective MOS had an average

correlation coefficient of 0.98 and 0.95 with refined MOS and coarse MOS respectively.

Further objective and subjective experiments were performed using a couple of other traditional African languages. The languages selected were isiZulu, isiSwana, isiXhosa (from Southern Africa) and Swahili and Dholuo (from eastern Africa). A total of 20 such sentences (10 male and 10 female) were used. The achieved correlation results are shown on table I.

Table I: Correlation of objective and subjective results before and after MOS Refinement.

Languages	Corr. before MOS refinement	Corr. after MOS refinement
isiZulu	0.95	0.98
isiSwana	0.96	0.99
isiXhosa	0.94	0.96
Swahili	0.96	0.99
Dholuo	0.97	0.99

In a similar manner, experimental measurements were run to evaluate the effectiveness of this new approach to the objective/subjective results correlation with test speech samples from different languages. A total of six languages were used in the experiments. The results are as shown in Figure 5 and 6. The samples were: (a) speech data from short sentences i.e. 10 seconds or less) and (b) speech data from longer sentences (i.e. 60 seconds or more) and results plotted. Most of the existing objective algorithms are tested with the English language samples first. Due to this, the English language was also used to enable a comparison of performance with relation to other objective estimators. Up to 40 different test measurements were run for each speech sample and the results computed. Figure 5 and 6 give a plot of these MOS outputs.

The graphs plotted in Figure 5 and 6 show that the objective quality outputs obtained by using African speech samples are consistent after repeated evaluation with many data samples. In figure 6, it was further observed that results obtained by using TCSQE measurement algorithm suffered minimal deviations among the different dialects used than the results from PESQ algorithm. This is attributed to the fact that TCSQE supports the test to be run using much longer speech samples (30 seconds or longer [8]) as opposed to what is recommended for PESQ (8seconds [3]).

In this study, it has been realised that the deviations that occur when using the same approach to evaluate speech quality on speech samples from different dialects is caused by the difference in phonetic structure of difference languages. Speech quality measurement systems should therefore incorporate other quality tracking data (i.e. metrics that define voice naturalness) for efficient design of accurate and reliable quality measurement algorithms for comparative QoS benchmarks. The incorporation of additional metrics that track voice naturalness is an important step towards developing a dialect-independent speech quality measure.

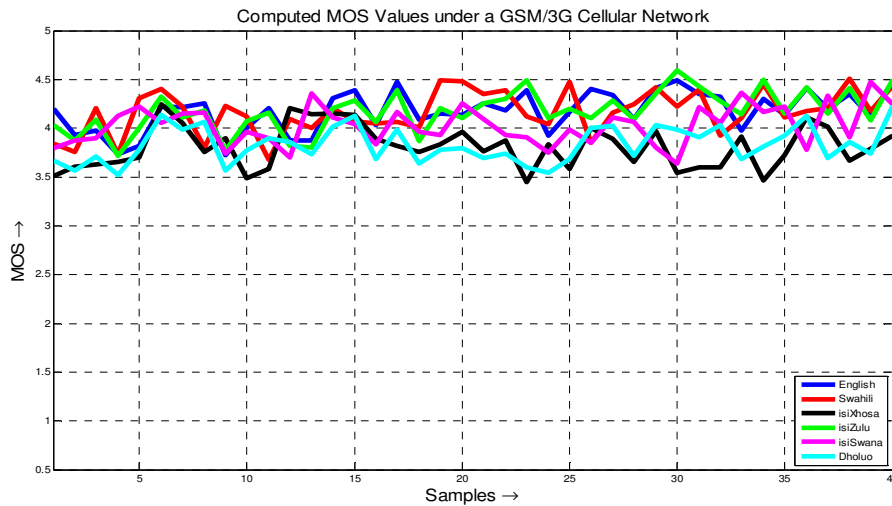


Figure 5: MOS Results of TCSQE with different languages under test data (a).

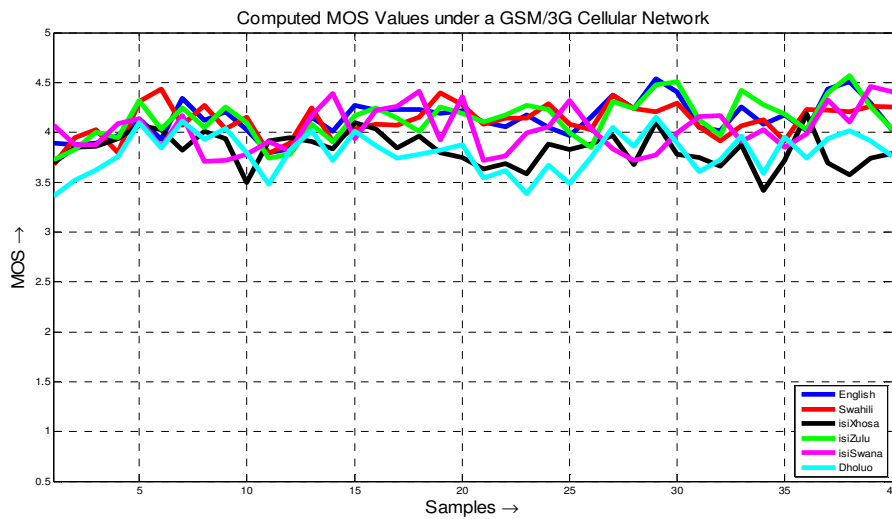


Figure 6: MOS Results of TCSQE with different languages under test data (b).

V. CONCLUSION

In this work, an algorithm for PESQ-MOS refinement has been proposed. The clustered parameter introduces the capability to detect talker dependency factors such as pitch and harmonic regions within natural speech samples. The parameters independently extract known attributes of natural speech from the original speech sample which are then compared with similar attributes extracted from the speech signal being tested.

Within the context of speech communications technology, selected areas of speech samples were investigated in more detail. The aim was to develop appropriate methodologies and metrics for auditory quality assessment and evaluation. In other words, while speech communications technology is the object of the study, the metrics for speech communication quality is the actual goal of investigation. The MOS refinement approach developed in this study is interdisciplinary since the topic requires knowledge in fields

as diverse as telecommunications engineering, human-perception research, linguistics, language technology and communications science.

The study and experiments were conducted using the context of quality of speech evaluation under cellular networks. However, it is a general technique that may be valid in many other situations as well, i.e. codec performance evaluation for example.

It has been established from the results achieved in section IV that MOS refinement gives a more accurate approach to estimating the quality of voiced speech transmitted through noisy cellular communication channels.

VI. RECOMMENDATIONS AND FUTURE WORK

In the most general case, speech/audio quality experiences using communication technologies is neither an absolute nor an inherent property of a telecommunications system, but depend on the specific users as well. Therefore, engineering approaches to quality of speech determination

should take consideration of how the speech are received/perceived by the users and how the listener understanding and expectations in communications develop with regard to naturalness of received speech.

With the development of remote access to information through various applications that depend on speech technology (speech recognition, dialogue management and speech synthesis), it is anticipated that MOS refinement will lead to accuracy of usage. To achieve this objective, further work is recommended to study the quality assessment of spoken-dialogue systems in real application scenarios.

It is further recommended for future research to concentrate on developing algorithms that analyse overall quality of communications. The evaluation methods that work on per sample approach do not fully reflect end-to-end user perception of speech quality because call duration is usually much longer than some seconds.

ACKNOWLEDGMENT

The authors of this paper are grateful for the support and guidance received from Mr. Owen Griffith of MTN-SA. Mr. Owen also provided the research materials which were used to perform live testing and experiments that made this work possible and successful.

REFERENCES

- [1] ITU-T Rec.P.800, "Methods for subjective determination of transmission quality", ITU-T, Geneva, August 1996.
- [2] ITU-T Rec.P.800.1, "Mean opinion score terminology". ITU-T, Geneva, March, 2003.
- [3] ITU-T Rec.P.862, "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", Geneva, February 2001.
- [4] S.Voran, "Objective estimation of perceived speech quality – Part II: evaluation of the measuring normalization block technique", IEEE Transactions on speech and audio processing, vol. 7, no. 4, pp.383-390, July 1999.
- [5] ITU-R Rec. BS.1387, "Method for objective measurements of perceived audio quality (PEAQ), ITU-T, Geneva, 1998.
- [6] ITU-T Rec. P.563. Perceptual non-intrusive single-sided speech quality measure, ITU-T, Geneva, 2004.
- [7] Adar, P.O, et al, "Optimal technique for speech quality evaluation on W-CDMA 3G cellular Networks". In *Proc. of 2007 IEEE 14th Int'l Conference on Telecommunications and 8th Malaysia Int'l Conference on Communications, ICT-MICC 2007*, Penang, Malaysia, May 14-17, 2007.
- [8] Adar P.O. 2008. Optimal Measurement of Speech Transmission Quality in GSM/3G Cellular Networks. Master thesis, Tshwane University of Technology, June 2008.
- [9] S. Voran, "Perception of temporal discontinuity impairments in coded speech – A proposal for objective estimators and some subjective test results" Proceedings of the 2nd international conference on measurement of speech and audio quality in networks, Prague, Czech Republic, May 2003.

- [10] KONDOZ, A.M. 1995. *Digital Speech: Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, New York, 1995.
- [11] J.G. Beerends, A.W. Rix, M.P. Hollier, and A.P. Hekstra "Perceptual evaluation of speech (PESQ), The new ITU-T standard for end-to-end speech quality assessment, Part II: Psychoacoustic Model," J. Audio Eng. Soc., vol. 50, no. 10, pp 755-764, oct. 2002.
- [12] Eric Larson, "Real time domain pitch tracking using wavelets. Available: www.online.physics.uiuc.edu/. [Accessed on : 24th July 2009]
- [13] Peter Mitev and Stefan Hadjitodorov, "A method for turbulent noise estimation in voiced signals", *Journal of medical and Biological engineering and computing*, Springer Berlin, Vol. 38(6), Nov.: 625-631.
- [14] FONG, L. C. 2005. Objective Speech Quality Measurement for Chinese Speech, *Msc Thesis*. University of Canterbury.
- [15] Anderson, J.: "Methods for Measuring Perceptual Speech Quality". Agilent Technologies white paper. USA, August, 2002.

P.O. Adar received his BSc in computer science and engineering from Maseno University (Kenya) in 2002, M-Tech degree in telecommunications technology from Tshwane University of Technology (Pretoria, South Africa) in 2008 and MSc in electronic engineering from ESIEE (Paris, France) in 2008.

He is currently working with Telkom Kenya Limited as a Solutions Consultant. His research interest includes quality of service in telecommunications networks, signal processing and broadband access technologies.